Be reasonable: a defence and development of the case for internal reasons.

Thesis

Be Reasonable: A Defence and Development of the Case for Internal Reasons

David Knott

PhD in Philosophy

2006

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

We see reasons as practical expressions of an agent's identity; we expect that if they govern and direct an agent's action they must be intimate products of his or her nature and circumstances. Yet we also believe that reasons are subject to external appraisal; agents can be wrong about their reasons, and external judgements concerning those reasons exert normative pressure. These thoughts are not necessarily incompatible while they remain informal intuitions. However, they are a source of conflict in philosophical theory.

I believe that the best way to satisfy the intuition that reasons are intimately ours is to follow the work of Bernard Williams, who claimed that only *internal reasons* exist: those which depend on the agent's motivations. In developing the case for internal reasons I attempt to show that it also goes some way to satisfying the intuition that our reasons are subject to external judgement, by allowing that we may be separated from our reasons by deliberative obstacles.

Admittedly, this satisfaction is only partial. Many theories attempt to show that reasons possess normative power precisely because they are not dependent on contingent factors such as motivations. I consider several such theories, particularly those which claim that aspects of human rationality determine the reasons of all agents. I attempt to show that they fail, partly due to specific flaws in their arguments, but also more generally because they assume that reasons must be universal.

The consideration of such arguments also helps develop the case for internal reasons. They reveal a common pattern of motivation and behaviour regarding reasons which is expressed in philosophical theory and in everyday talk. I argue that what is revealed is a virtue, and that, under the familiar name of reasonableness, this virtue provides the normative component our account needs, showing that the dependence of reasons on motivations is compatible with our common intuitions about reasons.

Table of Contents

1. The Argument for Internal Reasons

1.1 Reasons

Reasons matter. They matter because they express our rational nature, because they distinguish our actions from the sort of brute response to stimuli we might expect from an animal, a plant or even an inanimate object, and because, when we say that someone acted for a reason, we say something about choice and will. The existence of reasons is what makes it possible for us to ask the question, 'Why did you do that?' It is through our reasons that we justify and explain our actions to each other and to ourselves.

Furthermore, reasons do not just matter as abstract principles which remain in the background. We can stay silent about our reasons, but we rarely do. Our reasons for action, our thoughts about the reasons of others, and the deliberation which produces them, are all regarded as proper topics of everyday discussion. So, rather than keeping our judgements about the reasons of others to ourselves, we express them, and we do so with the intention of producing an effect, whether that effect is limited to informing others about our attitudes, or whether we hope to encourage or dissuade an agent embarking on a course of action. A challenge to someone to produce his or her reasons is likely to get a response. Even when the challenge is unexpected and we have no obligation to respond to the challenger, the thought that even a stranger might think that we are acting without good reasons is somehow discomfiting. Unlike many other philosophical preoccupations, such as the nature of knowledge, truth and the material world, the existence of reasons and the question of whether action is justified are subjects we talk about every day, albeit usually within the context of individual instances of action. Because talk of reasons is part of our everyday commerce with one another, we should take our everyday expectations and intuitions regarding reasons particularly seriously.

When we examine our everyday understanding of practical reason, I believe that we find an apparent conflict between two powerful intuitions. The first of these is that an agent's reasons for action are intimately bound up with the agent's concerns, commitments, affections, projects, relationships and other things that have a motivational component. This intuition highlights the appropriateness of the language we use when we say not just that there is a reason for A to ϕ but that A *has* a reason to ϕ. It is most apparent when, for example, we say of a man, 'He has his reasons,' meaning that there are reasons for his actions, even if few of us are likely to fathom them. This intuition emphasises the likelihood of difference between the reasons of agents, and acknowledges the possibility that exhortations to act for a particular reason might legitimately, if frustratingly, be met with the response, 'But I just don't care.' We can label this the *individualistic intuition*. At the same time, and in apparent contradiction, we have the intuition that external judgements about reasons ought to persuade agents and correct behaviour; that we can identify reasons which agents never knew they had, and that once these reasons have been pointed out to them they ought to have some bearing on their action. This intuition emphasises the ability of reasons to stand independently of the agent's ability or willingness to apprehend those reasons, and acknowledges that statements such as, 'But you've got every reason to ϕ!' have more than just rhetorical force. We can label this the *universalistic intuition*. The apparent conflict between the

individualistic intuition and the universalistic intuition has been reflected in philosophical debate. The individualistic intuition seems most closely compatible with theories such as that of David Hume, in which reason is taken to be subordinate to motivations, and it is allowed that reasons may vary between individuals as their motivations vary. The universalistic intuition seems most closely compatible with those theories in which rationality is taken to be the fundamental determinant of reasons, and consequently universal reasons are taken to apply to all agents regardless of their motivations.

It is not necessary to solve the apparent conflict between our everyday intuitions by showing that one is correct and the other mistaken, for we can achieve an understanding of practical reason which satisfies them both; or so I shall argue. However, this understanding is not achieved by staying neutral in the philosophical debate. I believe that the Humean position is the correct one, and that by taking into account recent developments of this position by Bernard Williams, as well as exploring and developing the position further ourselves, we can show both that it provides the best theoretical articulation and the best resolution of these two fundamental intuitions about reasons. In the first part of this thesis, then, I shall attempt to establish an essentially Humean position which not only shows how our reasons must be intimately ours, but also goes some way to satisfying the universalistic intuition by showing how our reasons can be obscure to us, and consequently why we may be subject to the judgements of external observers about our reasons.

However, I do not expect that this initial version of the account will fully satisfy the universalistic intuition. It will further emphasise the individuality of agents: their characters, their motivations, and even the way their reasoning happens to go in particular instances of deliberation. Consequently, it will need to answer challenges from those writers who seem to be driven predominantly by the universalistic intuition, and who favour an absolute conception of reasons as transcending the interests and motivations of the agent. I will attempt to show that the proposed account can resist such challenges and, furthermore, that such challenges depend on a conception of reasons which is not itself justified. I will also attempt to show that these challenges help us to understand the motivations underpinning the universalistic intuition, and that this understanding enables us to modify our account in a way which both better satisfies this intuition and stays true to its Humean origins.

These Humean origins are where we start our discussion.

1.2 Humean Origins

Hume's theories are primarily expressed in *A Treatise of Human Nature*, *An Enquiry Concerning Human Understanding*, and *An Enquiry Concerning the Principles of Morals*. I am not going to attempt to summarise all of Hume's work here. Much of it is concerned with topics which do not directly touch the matter of our discussion, such as causality and perception, and much that does deal with topics of interest to us comprises attempts to explain specific varieties of human behaviour, such as keeping promises, submitting to government and so on. This means that the elements of Hume's theory relevant to our discussion can be grasped by examining a relatively small number of key concepts and arguments. Before starting this examination, however, we should acknowledge a characteristic that pervades Hume's work. This is a distinctive blend of pessimism about the ability of reason to provide us with categorical foundations in the areas essential to our lives, combined with an optimism about our ability to go on with our lives nevertheless, and is best illustrated by an extended quote from the end of the first book of the *Treatise*, at a point when Hume takes himself to have shown that reason alone cannot support our beliefs in everyday phenomena such as causation and the persistent existence of objects:

> The *intense* view of these manifold contradictions and imperfections in human reason has so wrought upon me, and heated my brain, that I am ready to reject all belief and reasoning, and can look upon no opinion even as more probable or likely than another. Where am I or what? From what causes do I derive my existence, and to what condition shall I return? Whose favour shall I court, and whose anger must I dread? What beings surround me? I am confounded with all these questions, and begin to fancy myself in the most deplorable condition imaginable, inviron'd with the deepest darkness, and utterly depriv'd of the use of every member and faculty.
>
> Most fortuitously it happens, that since reason is incapable of dispelling these clouds, nature herself suffices to that purpose, and cures me of this philosophical melancholy and delirium, either by relaxing this bent of mind, or by some avocation, and lively impression of my senses, which obliterate all these chimeras. I dine, I play a game of backgammon, I converse and am merry with my friends; and when after three or four hour's amusement, I would return to these speculations, they appear so cold, and strain'd and ridiculous, that I cannot find in my heart to enter into them any further.[1]

This blend of optimism and pessimism introduces a tension which we will meet many times through our discussion: the tension between the possibility of using reason to seek more and more fundamental justifications for our beliefs and actions, and the ravenous tendency of reason to consume any justifications we offer without ever being satisfied. Hume's pessimism produces his conviction that a relentless use of reason will sweep away all foundations, and his optimism produces his conviction that we can nevertheless believe and act with justification. The elements of Hume's theory that are of direct interest to us can be loosely divided into those which express his pessimism by placing restrictions on the role of reason in influencing action, and those which express

[1] *A Treatise of Human Nature*, Book I, Part IV, Section VII, page 316.

his optimism by providing a constructive account of morals and virtue. We will consider the former now, as an introduction to our own basic position, and shall return to the latter much later.

To understand the restrictions Hume's theory places on reason we must have some understanding of his account of psychology, which begins with his taxonomy of the contents of the mind. In the very first sentence of the *Treatise*, Hume claims that, 'All the perceptions of the human mind resolve themselves into two distinct kinds, which I shall call IMPRESSIONS and IDEAS.'[2] At this stage Hume claims that the main distinction between ideas and impressions is in the 'degrees of force and liveliness, with which they strike upon the mind.'[3] It soon becomes apparent, however, that the most significant distinguishing characteristic of impressions is their primacy. The primacy of impressions is such that both the production and character of the experiences that constitute them are beyond our control. By contrast, ideas are secondary, as they are produced by reflection and imagination rather than direct experience, and, according to Hume, the character of their experience is limited to what has previously been experienced as impressions. The clearest example of the distinction between impressions and ideas is perhaps that between sensory perception and the memory of that perception. If I see a duck then I have an *impression* of a duck (a *complex* impression in Hume's terms, as my visual image of the duck is made up of lots of individual, *simple* impressions[4]). If I think about the duck after it has flown away then I have an *idea* of a duck. If the duck had feathers of a colour I had never experienced before I would not be able to imagine this colour until I had seen it; I would not be able to produce the idea without the preceding impression. To take another example from Hume, 'We cannot form to ourselves a just idea of the taste of a pineapple, without having actually tasted it.'[5] Although the case of sensory perception is the best way to illustrate the distinction between ideas and impressions, it also prompts obvious challenges. We could ask, for example, whether the direct perception of an object such as a duck really has primacy over our possession of the concept of a duck as a distinct object; contrast, for example, the experience of a casual visitor to a nature reserve who sees a duck, and a full time warden at that reserve, who sees not just a duck, but a particular species of duck, and possibly even a particular individual duck. When considering the experience of the warden we might wonder whether ideas precede impressions, or whether we ought to think in terms of precedence at all. Fortunately, as we are not directly concerned with the empiricist account of sensory perception, we can note the distinction between impressions and ideas within Hume's theory and move on to the next distinction which concerns us.

Just as the first book of the *Treatise* starts by making a distinction, so does the second: 'As all the perceptions of the mind may be divided into impressions and ideas so the impressions admit of another distinction into original and secondary.'[6] According to Hume, original impressions are those arising from direct sensory experience, while secondary impressions are those which 'proceed from some of these original ones, either

---

[2] *Treatise*, Book I, Part I, Section I, page 49.
[3] *Treatise*, Book I, Part I, Section I, page 49.
[4] *Treatise*, Book I, Part I, Section I, page 50.
[5] *Treatise*, Book I, Part I, Section I, page 49.
[6] *Treatise*, Book II, Part I, Section I, page 327.

immediately or by interposition of its idea.'[7] This distinction introduces the concept of the passions, which Hume classes as secondary impressions. Passions can best be understood as impressions whose content is sentimental rather than sensory, although they have the same primacy as sensory impressions. That is, their causes are not under our immediate conscious control, and their corresponding ideas cannot be formed without having experienced the passion first. We cannot successfully imagine what it is like to feel angry, for example, if we have never experienced anger. However, there is a crucial difference between sensory impressions and the passions, in that the passions are motivational; they are capable of leading to action. Hume offers an extensive taxonomy of the passions, identifying, 'pride, humility, ambition, vanity, love, hatred, envy, pity, malice, generosity,' as well as, 'desire, aversion, grief, joy, hope, fear, despair and security.'[8] However, he also argues that despite this array of passions, each of which has its own distinctive character, their motivational aspects reduce in the end to the expression they give to pain and pleasure: ''Tis easy to observe, that the passions, both direct and indirect are founded on pain and pleasure, and that in order to produce an affection of any kind, 'tis only requisite to present some good or evil. Upon the removal of pain and pleasure there immediately follows a removal of love and hatred, pride and humility, desire and aversion, and of most of our reflective and secondary impressions.'[9]

For us, the most important step comes when we combine the primacy of impressions, the status of the passions as impressions, and the ability of the passions to motivate. The implication of combining these elements is that the production of motivational states is analogous to the production of direct sensory experience. Of particular importance is that the production of impressions, whether sensory impressions or passions, has no necessary connection with reason. We do not use reason to discover what liver tastes like; that is simply the impression produced by eating liver. And we do not reason to the like or dislike of the taste of liver; the passion of like or dislike is simply the impression produced by the taste of liver. And it is this that leads to the expression of Hume's general scepticism about our ability to find foundations for action in reason in one of his most infamous and often quoted claims: 'Reason is and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them.'[10]

Unfortunately, this claim can sometimes be taken as the sum of what Hume has to say about the relationship between reason, action and the passions. As we shall see, there is rather more to his theory than the claim that reason is the slave of the passions. However, we should also note that even this provocative claim has rather more sophisticated implications than is sometimes imagined. In particular, we should note that, although it is expressed in a primarily negative form, it does not just deny a role to reason in the production of action, but allows a role as well. The role it denies is a motivational one, and is further expressed by Hume when he says, 'Since reason alone can never produce any action, or give rise to volition, I infer that the same faculty is as incapable of preventing volition, or of disputing the preference with any passion or action.'[11] The

---

[7] *Treatise*, Book II, Part I, Section I, page 327.
[8] *Treatise*, Book II, Part I, Section I, page 328.
[9] *Treatise*, Book II, Part III, Section IX, page 485.
[10] *Treatise*, Book II, Part III, Section III, page 462.
[11] *Treatise*, Book II, Part III, Section III, page 462.

point being made here is not that Hume expects passion to win in a battle with reason. When he observes that, 'Nothing is more usual in philosophy, and even in common life, than to talk of the combat of passion and reason, to give the preference to reason, and assert that men are only so far virtuous as they conform themselves to its dictates,'[12] he is not arguing that the preference ought to be given to the passions. Rather, he is arguing that reason and the passions are not even sufficiently similar kinds of entity as to come into conflict with one another. Within Hume's theory passions are original existences with their own distinctive experiential characters, whereas reason is 'the discovery of truth and falsehood.'[13]

The strength of Hume's conviction that passions and reason are distinct entities is expressed when he says, in another infamous passage, ''Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an *Indian* or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter.'[14] It is this sort of claim which can alarm people who hope, contrary to Hume, that reason is the master of the passions rather than the slave, and which creates the possibility of a bogeyman we shall meet later on: the *sensible knave*. Despite this fear, I do not think that anyone would seriously contest the claim that the passions – or whatever modern term we would use to label them – are capable of producing action. To do so would simply be to deny common experience. The challenge, then, in establishing a Humean or neo-Humean position, is not to show that the passions can produce action, but to demonstrate that the passions are *necessary* to produce action and that the action produced by passions can be understood as action for *reasons*. The picture we have painted so far, of passions which are produced with little control over their occurrence and no control over their character, yet are supposedly responsible for all rational action, seems at this stage to be thoroughly at odds with what we have called the universalistic intuition. Much of the rest of this thesis will attempt to address this concern, and there are already some elements present in Hume's theory which help us to do this.

The difference in kind between reason and the passions does not mean that there is no interaction between them at all. In contrast to the motivational role it denies to reason, the claim that reason is the slave of the passions explicitly allows reason an instrumental role. Slavery may be a subordinate relationship, but it is a relationship nonetheless. In its most basic form, and the form most straightforwardly implied by the claim that reason is the slave of the passions, the instrumental role of reason is to discover those courses of action by which passions can be satisfied. However, once we consider what is required to perform this role, we discover that the claim implies that reason can produce and dispel passions. Hume makes this implication clear when he says that, 'The moment we perceive the falsehood of any supposition, or the insufficiency of any means our passions yield to our reason without any opposition,' and, 'I may will the performance of certain actions as means of obtaining any desir'd good; but as my willing of these actions is only secondary, and founded on the supposition, that they are causes of the suppos'd effect; as soon as I discover the falsehood of that supposition, they must

---

[12] *Treatise*, Book II, Part III, Section III, page 460.
[13] *Treatise*, Book III, Part I, Section I, page 510.
[14] *Treatise*, Book II, Part III, Section III, page 463.

become indifferent to me.'[15] This idea, that reason can create or dispel passions, may seem to contradict the idea that reason is incapable of motivating, or that reason and the passions are so distinct as to be incapable of coming into conflict. At the very least it might seem to support the idea that even if the passions are required to produce action, if reason can produce passions, then action may ultimately be entirely governed by reason. This apparent contradiction can be resolved if we remember that Hume does not claim that reason plays no part at all in the guidance and production of action, but rather claims that reason *alone* cannot produce action. The point of the quotations just given is that if the agent possesses an original passion, the conclusions of reason excite or dispel passions with regard to actions insofar as those actions are shown to be useful or detrimental to the satisfaction of that original passion. Reason only excites or dispels passions by virtue of the motivational efficacy of the original passion, in a model of motivation which conforms to what R. Jay Wallace refers to as the 'desire-out, desire-in'[16] principle, even though he rejects the principle himself. To appropriate the language of another writer who rejects the Humean model of motivation, Thomas Nagel, we could say that, within such a model, it is reason which allows passions to 'transmit their influence over the relation between ends and means.'[17]

The existence of the instrumental relationship between reason and the passions also makes it possible for the passions to be justified by reference to reason, albeit in an extremely limited way. Hume argues that the only time a passion can be called 'unreasonable' is when it, 'is founded on the supposition or the existence of objects which really do not exist,' or when, 'we chuse means insufficient for the design'd end, and deceive ourselves in our judgment of causes and effects.'[18] In other words, passions can only be called unreasonable when they are based on these particular varieties of false belief, although even then, according to Hume, ''tis not the passion, properly speaking, which is unreasonable, but the judgment.'[19] This limited allowance for passions to be unreasonable to the extent that they are based on false beliefs is obviously consistent with a limited conception of reason as the discovery of truth and falsehood. However, the possibility that passions or judgements could be called unreasonable at all shows that, however limited, Hume's theory has room for a norm of rationality.

Furthermore, in addition to these basic allowances that reason has some authority, Hume's theory also contains three elements which he does not recognise as part of reason, but which operate in co-operation with reason, to the extent that they may be confused with it. Firstly, because he denies that reason alone is capable of discerning phenomena such as causation and the existence of the physical world which we nevertheless accept as part of our everyday lives, he allows that the gaps left by reason are filled by the imagination: 'So that upon the whole our reason neither does, nor is it possible it ever shou'd, upon any supposition, give us an assurance of the continu'd and distinct existence of body. That opinion must be entirely owing to the IMAGINATION.'[20] Secondly, he argues that the same operation of the imagination which leads us to believe

---

[15] *Treatise*, Book II, Part III, Section III, page 464.
[16] *How to Argue about Practical Reason*, in *Mind* (July 1990), page 370.
[17] *The Possibility of Altruism*, page 34.
[18] *Treatise*, Book II, Part III, Section III, page 463.
[19] *Treatise*, Book II, Part III, Section III, page 463.
[20] *Treatise*, Book I, Part IV, Section II, page 244.

in phenomena such as cause and effect also leads us to adopt general rules which, while they may not be based on evidence which is as consistent as that which implies causation, nevertheless informs our behaviour and judgements: 'Shou'd it be demanded why men form general rules, and allow them to influence their judgment, even contrary to present observation and experience, I shou'd reply, that in my opinion it proceeds from those very principles on which all judgments concerning causes and effect depend.'[21] Finally, Hume divides our passions into those which are calm and violent, allowing that not all of the passions which drive us seize us with the same urgency, but that some are barely perceptible. Now, Hume would be at pains to insist that these elements of his theory are not part of reason. In particular, with regard to the calm passions he warns that, 'Hence it proceeds, that every action of the mind, which operates with the same calmness and tranquillity, is confounded with reason by all those, who judge of things from the first view and appearance.' However, insistence that these are not part of the formal capacity of reason does not prevent us from including them within a full account of practical deliberation.

All of the aspects of Hume's theories of motivation, psychology and reason come together in his account of virtue to show just how far his theory stands from an understanding of human action as brute response to stimuli. We will return to consider Hume's account of virtue in more detail later, but for now will simply observe that it concerns common human patterns of motivation and behaviour, and our attitudes towards these patterns, and attempts to explain them by reference to human nature and inescapable aspects of our circumstances as well as the variable characteristics of culture. The resulting account allows that basic physical responses such as pain and pleasure are capable of producing complex and sophisticated behaviour such as action in accordance with the virtues of benevolence, justice, loyalty and respect for law. In summary, Hume gives us an account of motivation in which reason mediates passion with more or less sophisticated results. Furthermore, we do not have to accept every element of Hume's particular account of psychology to see that it is attractive, plausible and liveable.

However, for many this is not enough. The way in which Hume's account disappoints is perhaps best expressed by his allowance for the existence of a character with whom we shall become familiar: the sensible knave. Although Hume believes in common motivations, resulting in common reasons and common conceptions of virtue, this does not mean that the possession of these reasons and virtues is universal or necessary. So, on Hume's account, an agent who was not motivated by compassion or benevolence could ignore or even exploit the conventions and conceptions of virtue that others took as binding:

> And though it is allowed that, without a regard to property, no society could subsist; yet according to the imperfect ways in which human affairs are conducted, a sensible knave, in particular incidents, may think that an act of iniquity or infidelity will make a considerable addition to his fortune, without causing any considerable breach in the social union and confederacy. That *honesty is the best policy*, may be a good general rule, but is liable to many exceptions; and he, it may perhaps be thought, conducts himself with most wisdom, who observes the general rule, and takes advantage of all the exceptions.

[21] *Treatise*, Book I, Part III, Section XIII, page 197.

> I must confess that, if a man think that this reasoning much requires an answer, it will be difficult to find any which will appear to him appear satisfactory and convincing. If his heart not rebel against such pernicious maxims, if he feel no reluctance to the thoughts of villainy or baseness, he has indeed lost a considerable motive to virtue; and we may expect that his practice will answer to his speculation.[22]

In other words, the sensible knave may be a genuinely hopeless case, whose motivations do not provide the reasons for action which the rest of us share. Some people find this troubling, not just because the sensible knave is such an obviously reprehensible character, but because of the suspicion that if we allow that anyone does not share our reasons, we somehow exempt that person from judgement. In other words, although Hume's account seems fully attuned with the individualistic intuition about reasons, it seems at odds with the universalistic intuition; the reasons of the sensible knave are intimately his, but he apparently escapes those reasons which the rest of us want to ascribe to him. Fortunately, even before we have begun to develop our account beyond the position which Hume sets out in the *Treatise* and the *Enquiries*, we can see the beginnings of the ways in which we can allay the concerns of those people troubled by the sensible knave. Firstly, we do not have to restrict our judgements of the sensible knave to ascriptions of reasons; all sorts of other judgements are available to us, such as those of dishonesty or selfishness, and these seem to get closer to the heart of the matter. Secondly, and rather more importantly for our discussion, Hume has introduced the possibility, albeit in a limited way, that the sensible knave is mistaken about his reasons. As we have seen, Hume allows that judgements about actions taken to satisfy passions may be called unreasonable when they are based on errors of fact; on beliefs about objects which do not exist, or about means which do not satisfy ends. The possibility of going wrong in these limited ways may not seem to provide enough room to contain our misgivings about the sensible knave, but it at least raises the possibility that an apparent knave shares our reasons but, due to some obstruction in his reasoning, is cut off from them; that is, that the ascription of reasons from an external perspective is justified. The more our understanding of practical reason allows that reasons may be hidden from an agent, the greater its potential for satisfying the universalistic intuition in at least one way: by showing that our attempts to persuade others of reasons they deny may be more than just browbeating, and consequently may not be futile. Of course, this is only one way of satisfying this intuition, and is unlikely to be enough for those in whom the universalistic intuition is particularly strong.

With this thought in mind, and with the Humean context of our discussion established, we can now start to build the position we want to defend by turning to the arguments of Bernard Williams.

---

[22] *Enquiry Concerning Morals*, Section IX, Part II, 232.

1.3 Internal and External Reasons

The argument we are particularly interested in is expressed in two of Williams' essays, 'Internal and external reasons'[23] and 'Internal reasons and the obscurity of blame,'[24] and is further developed in his response to John McDowell's essay, 'Might there be external reasons?'[25] However, the themes of these essays also appear in Williams' other writings, particularly those which echo Hume's doubts about the power of purely rational enquiry.[26] Indeed, Williams acknowledges his debt to Hume by starting his argument with a position which he labels the sub-Humean model: '*A* has a reason to ϕ iff *A* has some desire the satisfaction of which will be served by his ϕ-ing.'[27] Of course, as is indicated by the label 'sub-Humean' this simplistic position is not exactly the one which Williams wishes to defend or explore, and neither does he imagine that it adequately reflects Hume's theory. As we have seen, although Hume claims that motivation is dependent on the passions, the passions are rather too complex to be simply labelled as desires, and the practical implications of humans possessing both complex passions and the capacity for instrumental reasoning are rather more profound than can be captured within the idea of the satisfaction of desires. However, even though the sub-Humean model is not directly espoused by Hume or Williams, it is nevertheless important. As well as providing a starting point, it encapsulates the thought that will be central to our account no matter how far we develop it: that practical reasons are somehow dependent on some aspect of our psychology, whether we call that aspect desire, passion, sentiment or motivation, which is itself not entirely subject to rational appraisal. The sub-Humean model is also important because, as we shall see, it is something like this model which writers who object to the idea that reasons are dependent on motivations often take themselves to be attacking, even though their criticisms may be irrelevant to more sophisticated accounts such as those of Hume and Williams.

Williams takes us the first step beyond the sub-Humean model not by talking directly about reasons, but by talking about what we mean when we make statements about reasons. He claims that when we ascribe a reason to an agent – for example, by saying, 'You a have a reason to ϕ,' – we can intend this statement in one of two distinct ways. We may intend it as a statement whose truth is dependent on what Williams calls the agent's *subjective motivational set*.[28] Or we may intend it as a statement whose truth is independent of the contents of the agent's motivational set. So, to say, 'Go on, buy it and treat yourself. You really want it, and that's reason enough,' is to make a statement of the first sort, while to say, 'It doesn't matter if you've stopped caring about winning or the future of the team. Just get out there and play!' is to make a statement of the second sort.[29] Drawing a distinction which we shall follow throughout this thesis, Williams calls

---

[23] 'Internal and external reasons' in *Moral Luck.*

[24] 'Internal reasons and the obscurity of blame' in *Making Sense of Humanity.*

[25] 'Might there be external reasons?' in *World, Mind and Ethics.*

[26] For example, see *Ethics and the Limits of Philosophy*, especially Chapter 10: 'Morality, the peculiar institution', and the essays 'Persons, character and morality' and 'Moral luck' in *Moral Luck.*

[27] 'Internal and external reasons', page 101.

[28] 'Internal and external reasons', page 102

[29] Or, more properly, may be to make a statement of the second sort. As we shall see, one of the complexities of this discussion is that statements about reasons may appeal to more obscure motivations than the ones they explicitly mention.

those statements whose truth is dependent on the contents of the agent's motivational set *internal reason statements*, and those statements whose truth is not dependent on the contents of the agent's motivational set *external reason statements*. The central claim in Williams' argument – his equivalent to the thought encapsulated in the sub-Humean model – is that there are no true external reason statements. That is, any claim that a reason exists regardless of the contents of the agent's motivational set is false. This does not mean that the ascriptions of reasons contained in external reason statements are necessarily false; I may have a reason to play for my team because I want to impress somebody in the crowd or because I want to get the coach to stop haranguing me. What is false is the implication that even if my motivational set did not contain anything which provided such a relation to a reason to play, then the reason would persist.

We may ask why, if external reason statements are never true, anybody would ever make them. There are three immediate explanations, each of which supposes that the supposed external reason statements are not actually intended as such statements after all. Firstly, we may imagine that an apparent external reason statement is simply a mistake. Someone is insisting that an agent has a particular practical reason, but, through a thorough misunderstanding of his or her motivations, is just wrong. For example, my friend insists that I should try his home made ice-cream because I will like it, and I continuously refuse because I have a tooth cavity that makes eating any cold substance excruciatingly painful, but am too embarrassed to say so.

The second explanation is similar but rather more general. The person making an apparent external reason statement is actually making what Williams calls an *optimistic internal reason statement.*[30] Such a statement may appear to be about external reasons, because it takes little account of variation in the psychology of individual agents, but actually concerns internal reasons because the person making the statement hopes that we all possess the motivations which lead us to act in accordance with the statement. So, someone who says that we all have a reason to give to charity may seem to be claiming that this reason exists regardless of our individual feelings of compassion or benevolence, but may rather be hoping that all of us possess, however well hidden, the motivations necessary to support this reason.

The third explanation for making external reason statements is rather more sinister. Williams suggests that many such statements are bluff[31]: that they are not intended seriously to persuade the agent that he or she has the reasons claimed in the statement, but are rather intended to *give* the agent reasons to act in accordance with the statement, by appealing to existing motivations such as the fear of being regarded as irrational. So, an unscrupulous salesman might say, 'I don't understand why you're hesitating – you've got every reason to buy this product,' not because I do have every reason to do so, but because he hopes that invoking reason has the power to persuade me. I think that while Williams' identification of the phenomenon of bluff is insightful, neither it nor the other explanations we have considered exhaust the reasons why people make external reason statements. Sometimes they do not just want agents to act in accordance with the reasons which they claim they possess: they want them to actually have those reasons, and they want those reasons to transcend the accidents of the individual psychologies of agents. We will return to these motivations concerning

[30] 'Internal reasons and the obscurity of blame' in *Making Sense of Humanity*, page 40.
[31] See 'Internal and external reasons', page 111.

external reasons much later, in the final part of this thesis.

The claims made within Williams' argument obviously require justification. However, before attempting to find such justification, we will pause for a moment to get our terminology clear. This is especially important as the terms 'internal' and 'external' are used in different ways in various areas of philosophy, some of which touch directly on our discussion. As we have seen, Williams talks about internal and external reason *statements*. However, as it is clumsy to continuously talk of statements, and as our main concern is with reasons, we will talk of internal and external *reasons*. We will also use the distinction between different types of reasons to distinguish between different types of writers. So, we will use the term *internal reasons theorists* to refer to those writers who maintain, with Williams and Hume, that reasons are dependent on motivations, and will use the term *external reasons theorists* to refer to those writers who maintain that reasons can exist independently of motivations.

Because we will often be talking of moral reasons and of motivation, we also need to distinguish our debate about internal and external reasons from the debate concerning internalism and externalism about moral motivation. This debate concerns the question of whether espousing a moral belief entails the possession of a motivation to act in accordance with that belief. The distinction is often first credited to W.D.Falk[32] and a version of it is expressed by Thomas Nagel when he says that, 'Internalism is the view that the presence of a motivation for acting morally is guaranteed by the truth of ethical propositions themselves,' and that, 'Externalism holds, on the other hand, that the necessary motivation is not supplied by ethical principles and judgements themselves, and that an additional psychological sanction is required to motivate our compliance.'[33] Internalists about moral motivation need not be internal reasons theorists and external reasons theorists need not be externalists about moral motivation: one could argue that believed ethical propositions produce motivation in the absence of pre-existing motivations, or that believed ethical propositions could fail to motivate just because of a lack of prior motivations. Indeed, Nagel declares himself to be an internalist about moral motivation but, as we shall see, argues strongly for the existence of external reasons.[34]

As we are setting the terms of our discussion, we should also clarify how we will speak about the things that, on Williams' and Hume's accounts, give rise to reasons. We have already seen that there are several ways to refer to these entities, and we shall see several more. The term chosen by particular writers reflects the language of their times, but also reflects their theoretical positions. The term 'passion' does not have quite the same meaning to us as it had in the 18th century,[35] but I expect that in both times it implies that what it refers to provides the impetus to action. Williams, who is concerned to make sure that we do not underestimate the range or complexity of motivations, deliberately adopts the technical term 'subjective motivational set' and gives it the

---

[32] See "'Ought' and Motivation' in *Ought, Reasons and Morality.*

[33] *The Possibility of Altruism*, page 7.

[34] The possibility for confusion between the terms is shown by Christine Korsgaard when, in her paper 'Skepticism about practical reason' she claims that Williams' and Nagel's understanding of internalism is 'almost identical.' As we shall see, they are thoroughly at odds. 'Skepticism about practical reason' in *Creating the Kingdom of Ends*, page 329.

[35] Or, indeed, the wide range of meanings it apparently had in the 17th century. For an extensive discussion of the treatment of the passions in this period, see Susan James' *Passion and Action: The Emotions in Seventeenth-Century Philosophy.*

symbol *S*, although he recognises this as 'unlovely.'[36] As we shall see, writers who are sceptical about the role that non-rational psychological entities can play in the determination of reasons, tend to talk of desires, a term which carries implications of crudity and irrationality. For the sake of simplicity I shall mostly talk of motivations, although I shall also adopt the language of whatever writer we are discussing at the time.

The final aspect of terminology which we must settle is what we will call the account of practical reason defended in this thesis. I shall call it the *internal reasons account*, although it is not exactly the same as that which appears in Williams' essays, nor is it exactly the same as that in Hume's work. Furthermore, it is deliberately labelled an account rather than a system or a theory, both of which imply a degree of unity and comprehensiveness which I do not attempt to attain. The account is not complete, and although it will be more complete at the end of this thesis than it is now, it will never be a thoroughly systematic, unified whole. Rather it will be sufficiently sophisticated to be plausible, recognisable and resilient; and, I shall attempt to show, more so than any other alternatives we encounter.

Given this ambition, it is time to see how the claims about internal and external reasons we have adopted from Williams' account can be justified. Williams' own justification of his claim is quite terse: 'If there are reasons for action, it must be that people sometimes act for those reasons, and if they do, their reasons must figure in some correct explanation of their action.'[37] This claim implies much but its terseness is regrettable, as it can lead to simplistic interpretations. I believe that there are two complementary legitimate interpretations, and that they reflect two inescapably linked dimensions of reasons. The first interpretation is that it is a condition of practical reasons that they constitute forces capable of driving agents into action, so they therefore must contain a motivational component. This interpretation reflects the explanatory dimension of reasons; the thought that when we say that something was done for a reason, we can point to the reason as part of the explanation. This expectation that reasons can explain is undoubtedly part of our understanding of reasons, but it is not the only part. That it is not the only part is indicated by our tendency to use the language of reasons when explaining the behaviour of inanimate objects: 'The reason for the avalanche was the heavy fall of snow last night.' When talking of the reasons of rational agents we must say something more. This something more comes in the second interpretation of the claim that if there are reasons for action, it must be possible that agents could act for them. This interpretation points out the normative implications of the claim: the ascription of a reason constitutes a judgement, and to judge that an agent *should* act for a reason which he or she *could not* act for cannot be justified. This idea is often expressed as the principle that 'ought' implies 'can'. So, leaving aside the question of motivation, a judgement that someone should stop a runaway car from rolling into a crowd of people is unjustified if it is physically impossible for the car to be stopped. Similarly, if a particular motivation is a pre-requisite for a particular action, then the absence of that motivation is as much an obstacle to that action as any physical restraint. It is important to realise just how tightly these two interpretations, and consequently the explanatory and normative dimensions of practical reason, are bound up with one another. The possibility of explanation justifies normative judgement, while the judgement of normativity is what

[36] 'Internal reasons and the obscurity of blame', page 35.
[37] 'Internal and external reasons', page 102.

elevates the ascription of reasons to rational agents above the explanation of the behaviour of inanimate objects. The normative judgement and the motivational impetus are both necessary to provide a reason.

It seems, then, that, despite Williams' terse treatment of the subject, we can justify the claim that motivations within an agent are necessary for the existence of reasons for that agent; motivations underpin both normative and explanatory aspects of reasons. However, I think that there is also a less formal basis for the claim embedded within the internal reasons account; and that is the individualistic intuition about reasons which we introduced at the beginning of our discussion. Normativity and explanation are essential characteristics of practical reasons, but it is also an essential part of our common understanding of practical reasons that they are intimate to the agent, and that those reasons which are most important to the agent are those that lie closest to the essential elements of his or her identity. This is why, even though we expect that we will share many of our reasons with others, I do not expect that my reasons will be exactly the same as yours; and I expect that this difference is, at root, a function not just of the circumstances in which I find myself, but of who I am. So, the claim that reasons are dependent on motivations, an aspect of psychology which is thoroughly individual, not only captures the normative and explanatory aspects of reasons, but also satisfies our intuition that some of the most important reasons to agents are the ones which are most individually theirs.

However, as with Hume, there is more to Williams' argument than the basic conditions for the existence of reasons. Hume went beyond his account of motivational psychology to provide accounts of morals, virtues, social conventions and their possible historical origins. Williams does not explicitly develop his account in this way, but does argue for three additional claims which are vital to his account: that we should have a broad and generous understanding of the deliberation that produces reasons; that we should have a similarly broad and generous understanding of those things which comprise our subjective motivational sets; and that the deliberative routes from our existing motivations to our reasons, and consequently our reasons themselves, are not fully determinate.

Williams' most succinct expression of his position is as follows: 'A has a reason to ϕ only if he could reach a conclusion to ϕ by a sound deliberative route from the motivations he already has.'[38] The obvious question raised by this formulation is what constitutes a sound deliberative route. In 'Internal and external reasons' Williams says that although means-end, instrumental reasoning is the most obvious type of sound practical deliberation, it is by no means that only one:

> But there are much wider possibilities for deliberation, such as thinking how the satisfaction of elements in S can be combined, e.g. by time-ordering; where there is some irresoluble conflict among the elements of S, considering which one attaches most weight to (which, importantly, does not imply that there is some one commodity of which they provide varying amount); or, again, finding constitutive solutions, such as deciding what would make for an entertaining evening, granted that one wants entertainment.[39]

[38] 'Internal reasons and the obscurity of blame', page 35.
[39] 'Internal and external reasons', page 101.

And in 'Internal reasons and the obscurity of blame', he points out again that a sound deliberative route is not just the identification of means to ends:

> There are many other possibilities, such as finding a specific form for a project that has been adopted in unspecific terms. Another possibility lies in the invention of alternatives. One of the most important things deliberation does, rather than thinking of means to a fixed end, is to think of another line of conduct altogether, as when someone succeeds in breaking out of a dilemma. Yet another line of deliberative thought lies in the perception of unexpected similarities.[40]

This broad understanding of what constitutes a sound deliberative route means that we should not allow ourselves to get caught up in the somewhat limited debate about whether, by claiming that reasons are dependent on motivations, we are implying that we can reason about means but not about ends. Our account of internal reasons does imply that our ends are dependent on our motivations, but still allows that we can reason about these ends: our pre-existing motivations constrain our ends but they do not determine them.

As with the concept of sound deliberation the importance of the concept of the agent's motivational set within Williams' theory is not just that it determines the existence of reasons, but that it is understood broadly. Williams is as wary of the term 'desires' as we are, noting that although it is convenient and familiar, 'this terminology may make one forget that S can contain such things as dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may be abstractly called, embodying commitments of the agent.'[41] Furthermore, the contents of S are not fixed but dynamic. As Williams points out, 'The process of deliberation can have all sorts of effect on S, and this is a fact which a theory of internal reasons should be very happy to accommodate.'[42] Some of these effects may be of the particularly rational kind acknowledged by Hume, such as the instrumental desire for an object imagined to satisfy a more fundamental desire vanishing when it is realised that the object doesn't satisfy that desire at all. But some of these effects may be less predictable and less directly rational, such as an upwelling of enthusiasm caused by imagining a particular goal or course of action.

It is this broad understanding of motivations and deliberation and the interaction between them which makes Williams' account more than just the claim that external reasons do not exist, and which raises the possibility of an account of internal reasons which goes far beyond the sub-Humean model, but which retains the plausible and attractive aspects of Humean motivational psychology. Such an account seems even more possible when we consider the last additional element of Williams' account: indeterminacy. The broad understanding of motivations and deliberation means that Williams emphatically does not imagine that, if we start with a comprehensive understanding of an agent's motivations and of what constitutes sound deliberation we can lay out the agent's reasons like a map. Rather, Williams allows that, 'There is an

---

[40] 'Internal reasons and the obscurity of blame' page 38.
[41] 'Internal and external reasons', page 105.
[42] 'Internal and external reasons', page 105.

essential indeterminacy in what can be counted as a rational deliberative process,'[43] and that, 'Since there are many ways of deliberative thinking, it is not fully determinate in general, even for a given agent at a given time, what may count as 'a sound deliberative route'; and from this it follows that the question of what the agent has reason to do is itself not fully determinate.'[44] And this indeterminacy accords with our general experience of reasons and reasoning: we often feel that there is not a single, compelling reason for taking a particular action at a particular time, but rather an almost infinite range of potential reasons and actions without any single right answer. It is a simultaneously luxurious and disconcerting feature of life for many of us fortunate enough to have lives which are not dominated by the practicalities of survival that this is more often the situation than not.

However, while we can allow and even welcome a certain indeterminacy in reasons, we should try to be rather clearer about what we mean by indeterminacy. I think that we can understand it best by considering three of its more important implications. Firstly, when we allow that what a particular agent has reason to do is indeterminate we allow that it need not be contrary to reason for agents with identical starting motivations and circumstances to deliberate to different and possibly contradictory conclusions about action, and yet for their reasons to be both justifiable and appropriate to them. This is possible because, once we acknowledge that imagination is an essential part of deliberation, we accept that reasoning can go in many different directions from the same starting point. Furthermore, once we allow that deliberation can modify motivations, we create the possibility that deliberation can be *one way*; that is, deliberation can modify motivations in a way which means that the original starting point and the deliberative routes away from it are no longer available to the agent. It may be helpful to illustrate this with an example. Imagine twin brothers in a country occupied by an invading force, both of whom are deliberating about whether to join the resistance or to try to survive within the constraints imposed by the occupier. For the purposes of this example we will imagine that they are identical twins to an implausible degree, with the same sets of motivations, circumstances and access to information. So, they are both proud of their country and hate the invader, but also both know the retribution exacted on resistance fighters and their families if caught. While deliberating, one of them considers the proud tradition of his family in similar conflicts, the indignities he would suffer under occupation, and the respect afforded the resisters. He is fired with patriotism and dreams of heroism, and comes to regard anything but resistance as despicable capitulation. By contrast, the other considers his love for his family, their certain torture and death if he was caught, and the futility of a few fighters in the face of the invading forces. He becomes depressed and resigned, seeing anything other than survival, with all the compromises that entails, as foolhardy posturing. Both end up with different reasons from the same starting point, and once they have reached those reasons, no longer have the same motivational set as the other. The obvious question is what could cause such divergence from the same starting point. We could work somewhat implausible triggers into the example, imagining these two brothers pondering side by side, one whose gaze falls on his father's old rifle, and the other whose gaze falls on a picture of his family. However, I think that it is sufficient to acknowledge that, once we accept a broad

[43] 'Internal and external reasons', page 110.
[44] 'Internal reasons and the obscurity of blame', page 38.

understanding of deliberation which includes the operation of the imagination, there is no set way that deliberation must go in particular circumstances; there are many ways to go which count as sound, and many of these are paths which cannot be retraced once taken.

The second important implication of indeterminacy is that we cannot determine exactly what an agent's reasons are before that agent deliberates, because the existence of those reasons depends on the process of deliberation as well as the point from which deliberation starts. The example of the twin brothers living in the occupied country is useful once again here. Our first illustration of indeterminacy allowed that the two brothers could settle on contradictory reasons from the same starting point. Our second implication of indeterminacy is that even with a thorough understanding of the psychology of each brother, until actual deliberation is conducted and actual motivations are influenced, we cannot know or say that the reasons to join the resistance or to avoid conflict definitely exist. This does not mean that we cannot know any of the agent's reasons; if we allow that motivations give rise to reasons then some immediate motivations will directly produce some reasons, and the predictability of reasons will vary with the degree and complexity of deliberation required to produce them. However, at some stage we will reach the point beyond which the outcome of deliberation cannot be known prior to deliberation. We will return to this thought later, and particularly to the question of what precedes the motivations which emerge during deliberation.

The third implication of indeterminacy is perhaps the most important, as it concerns our nature as reasoning agents, rather than just the questions of how our reasons can diverge or what we can know about each others' reasons. The allowance that our reasons are not fully determined prior to deliberation provides us with a sense in which we can be both rational and free. As we shall see later, some theorists, especially those who follow Kant, have attempted to show that rational beings are free precisely because their status as rational agents fully determines their reasons. Leaving such arguments aside for the time being, however, the idea that reasons are fully determined, whether we suppose that those determinate reasons are universal or particular to individual agents, implies that when we deliberate correctly we do not *decide* what to do, but rather *discover* those determinate reasons which were in place all along: it seems to leave us with no room for choice, or at most the choice between rationality and irrationality. By contrast, once we allow that reasons are not fully determined, we allow that when we are deliberating we are doing more than navigating a landscape which has already been laid out for us, but are involved in real choices in which we construct our reasons as well as discover them.

To some it may seem that if we not only allow indeterminacy, but positively embrace it, we are constructing a vision of chaos rather than of practical rationality. If deliberation can go in any direction, and if we cannot tell what reasons an agent may end up with, it may seem that we are no longer talking about anything recognisable as reasons at all. Of course, though, Williams is not proposing anarchy. Indeterminacy is only a part of the internal reasons account, even if it is an important part. Indeed, within his argument, Williams insists that reasons are subject to at least the same minimum standards of rationality that Hume imposed on whether a passion could be considered reasonable or not. As we saw, Hume allowed that passions which were based on certain types of false belief could be described as unreasonable. Williams expands on this thought by offering the example of a man who wants gin and is unaware that the gin

bottle in front of him actually contains petrol.[45] When we ask whether the man has a reason to drink what is in the bottle, Williams acknowledges that such action could be explained by reference to the agent's false belief, but maintains that, precisely because the action would be based on a false belief, the agent does not have a reason to act that way.[46] In other words, even if we allow that reasons are indeterminate, and that what counts as a sound deliberative route is indeterminate, we do not allow that just anything could count as a reason or as a sound deliberative route. Allowing that reasons are not fully determined is not the same as allowing that they are arbitrary.

A picture of the deliberating agent is beginning to emerge. It is one in which the agent possesses a starting set of motivations, and a set of sound ways of proceeding from those motivations through deliberation. These ways of proceeding will produce practical conclusions which are capable of explaining and justifying actions. The agent does not necessarily have to follow all these paths for all these reasons to exist, but it has to be conceivable that they could be followed. The attraction of this picture is not only that it is plausible and recognisable and that it has room both rationality and freedom, but that, at least at first glance, it appears comprehensive and resilient. It can be used to account for basic actions, peculiar to the desires of a particular individual, but can also be used in the hands of someone like Hume to construct a grand and general scheme of human behaviour. It does not seem as if there is much we need to add to the fundamentals of the account to encompass all of our motivations, actions and reasons.

As in our discussion of Hume's scheme, though, we must acknowledge that Williams' account of internal reasons seems, at least superficially, rather more agreeable to the individualistic intuition about reasons than it does to the universalistic intuition; the emphasis on motivations and the likelihood that external reason statements are mere bluff seems more compatible with the thought that our reasons are intimately ours than with the thought that we can legitimately make judgements about the reasons of others from an external standpoint. However, also as with Hume, Williams' account contains some elements which indicate how it could potentially satisfy the appetite for objectivity underlying the universalistic intuition. The idea that deliberation must be sound implies that there is some standard, albeit an undefined and indeterminate one, which constrains what can be allowed to count as a reason. Furthermore, the idea that reasons are not always produced directly by motivations, but may lie at the end of a sound deliberative route which is long and complex and which itself involves the modification of the agent's starting motivations, allows that agents may possess reasons which are obscure to them; indeed, most of us almost certainly do possess such reasons. Consequently, exhortations to agents to pay attention to reasons which they are apparently unaware of or indifferent to may often be more than just bluff or wishful thinking; they may be attempts to get agents to see reasons which they possess but which they nevertheless cannot grasp without help. As mentioned earlier, the more our account allows for the obscurity of reasons, the more it allows for the possibility that our judgements about an agent's reasons may be correct, even though they disagree with that agent's own judgements about his or her reasons. To take this thought further, we will develop Williams' account in two directions. Firstly, we will show that the account allows for an even greater degree of indeterminacy than we have acknowledged so far, by considering the implications of

[45] See 'Internal and external reasons', page 102.
[46] See 'Internal and external reasons', page 102.

ideas presented in Williams' recent book *Truth and Truthfulness.* Secondly, we will further consider the implications of the possibility that agents may find their deliberative routes obstructed, and what this means for their reasons.

1.4 Steadying the Mind

Although the idea of a starting set of motivations figures formally in Williams' account of internal reasons to the extent that it is given the label *S*, it would be a mistake to imagine this starting set of motivations as stable, well-defined and transparent. When we consider our experience of motivations, we realise that they are often highly unstable, and, furthermore, that we are often unsure or unaware of what our motivations are until we are called upon to deliberate in some way which engages them. Consider, for example, the percentage of 'don't knows' in virtually any survey which is conducted. Some of these people undoubtedly say they don't know which option they prefer because they don't like any of them, but others genuinely don't know what their motivations are, and could even be said to have no motivations relating to the question until after they have been asked. This seems to be the case whether the choice being offered is between politicians or bars of chocolate. The phenomenon of the contents of agents' motivational sets being clarified in deliberation was discussed by Williams in his work on internal and external reasons. However, he also considered it much more recently in *Truth and Truthfulness*, in which he introduces the useful concept of *steadying the mind.*

Williams introduces this concept by considering two different but equally unlikely characters from 18th century French literature. These are both historical figures, although one is presented via autobiography and the other is presented as a fictionalised caricature. The autobiographical figure is Jean-Jacques Rousseau as he presents himself in *The Confessions*, and the caricature is Jean-François Rameau as he is presented in Denis Diderot's dialogue *Rameau's Nephew*. Superficially, it is remarkable how much these two characters resemble each other, or at least how much Rameau resembles Rousseau as he presents himself in the earlier part of *The Confessions*. Both are moderately well educated but penniless, both depend on patronage for survival, and both attach themselves to households where they are found to be companionable and entertaining, but add little of any real substance. Inveterate hangers-on, both have earned their livings at one time or another as rather indifferent music teachers, and there is even an episode in *The Confessions* in which Rousseau could be following exactly the advice of Rameau on how to bluff a living out of teaching music without really understanding the subject at all.[47] The main difference between the two characters is the apparent conception of the self under which they are operating.

At the outset of *The Confessions* Rousseau famously declares, 'I desire to set before my fellows the likeness of a man in all the truth of nature, and that man myself,'[48] and it is apparent from this declaration and much of the rest of the book that he believes that there is a core self, an essential Rousseau to present. Moreover, it also becomes apparent that, even though Rousseau claims to 'have neither omitted anything bad, not interpolated anything good,'[49] and some of what he refuses to omit includes venal, cowardly, selfish and lecherous acts, he does not present himself as essentially vicious. Williams points out that Rousseau has a tendency to report something terrible and then to find some means of, if not exactly forgiving himself, at least offering himself some form of release from responsibility. What is of most interest about such incidents is not

---

[47] See *The Confessions*, pages 142-145 and *Rameau's Nephew*, pages 58-60.
[48] *The Confessions*, page 3.
[49] *The Confessions*, page 3.

whether Rousseau is truthful or sincere in his account of them, but that he does not treat them as revelations of his essential self, regarding them rather as unfortunate moments of weakness. As the catalogue of deceit, cowardice and betrayal grows, however, we may begin to suspect that, despite all his apparently sincere self-recrimination, Rousseau can only bring himself to recount such painful events because he does not take them so much to reveal his character, as to be occasions on which he acted out of character.

By contrast, the character of Rameau as portrayed by Diderot cheerfully represents himself as fickle and subject to vice. However, it would be too simple to write him off as a cynical egotist. There is an element of cynicism to Rameau, as revealed when he complains to his counterpart in the dialogue (whom we take to be Diderot himself) of the drudgery that flattery, the procurement of patronage, and promoting the interests of his patrons entails.[50] However, he is most interesting because he is not straightforwardly cynical or hypocritical, but rather unsteady to an extreme in his beliefs, his feelings and even his personality. It does not seem to simply be the case that he wears the mask that pleases his patron of the moment, but rather that he fully inhabits whatever role he is playing at the time. Furthermore, he is conscious of his unsteadiness:

> 'Devil take me if I really know what I am. As a rule my mind is as true as a sphere and my character as honest as the day: never false if I have the slightest interest in being true, never true if I have the slightest interest in being false. I say things as they come to me; if sensible all to the good, but if outrageous, people don't take any notice. I use freedom of speech for all it's worth. I have never reflected in my life, either before speaking, during speech or after. And so I give no offence.'[51]

His protean nature is even exhibited in his physical appearance. Diderot says that, 'Nothing is less like him than himself. At times he is thin and gaunt like somebody in the last stages of consumption . . . A month later he is sleek and plump as though he had never left some millionaire's table or had been shut up in a Cistercian house.'[52] Such extraordinary changeability might lead us to wonder whether this character possesses any real identity at all; whether there is anything behind the façade. Yet Diderot presents Rameau as such a vivid character with such a palpable presence that the situation is evidently the other way around: what Rameau feels and believes at any time goes to his core; there is just no guarantee that he will believe or feel it a day or even a moment later. Nothing sticks.

Although we can recognise aspects of ourselves in Rousseau and Rameau, they are ultimately both implausible: the products of dramatic invention or self-deception.[53] And it is what makes them implausible that helps us to understand to what degree we expect the contents of our minds to be settled, and what is going on as they become steadied. Rousseau is implausible because he asks us to believe that he possesses a fundamental character to which he is unswervingly true, even when he is behaving in

---

[50] See *Rameau's Nephew*, pages 87-88.
[51] *Rameau's Nephew*, page 79.
[52] *Rameau's Nephew*, page 34.
[53] As Hume said of Rousseau, 'I believe that he intends seriously to draw his own picture in its true colours: but I believe at the same time that nobody knows himself less.' Quoted in *Truth and Truthfulness*, page 177.

what appears to be complete contradiction to this character. While we feel there are elements of our characters which are fundamental to us, we do not suppose that they are so extensive as to inform all of our behaviour and beliefs, or so robust that they can survive behaviour which is consistently at odds with them. By contrast, Rameau is implausible because he transforms himself so thoroughly to suit his circumstances. While we do not necessarily believe that our characters can be so firmly established and maintained as Rousseau believes his to be, we do believe that we have characters that possess a persistent identity and inertia, placing Rameau's level of volatility beyond us. We are not like Rameau, because we need at least some stable parts of our characters to operate as agents at all, and we are not like Rousseau as he presents himself, because elements of our character are unstable, and are influenced by the roles we adopt and the acts we perform.

However, Williams evidently believes that we are rather more like the character of Rameau than we are like Rousseau as he believes himself to be. The idea that each of us possesses a thoroughly steady character which is transparent to us and to which it is incumbent upon us to be true is taken to be an illusion, at least partly an artefact of the artificial and rather romantic notion of authenticity: 'So we must leave behind the assumption that we first and immediately have a transparent self-understanding and then go on either to give other people a sincere revelation of our belief from which they understand us (or, as Rousseau bitterly found out, misunderstand us), or else dissimulate in a way that will mislead them.'[54] Rather, Williams maintains that, throughout our lives, whether through influences extended over time, such as upbringing or interaction with the society in which we live as mature adults, or through more acute turning points, such as direct engagement with individuals or particular instances of deliberation, we steady the contents of our minds, settling barely formed thoughts and feelings into the more determinate stuff of identity, knowledge and action. We are not as volatile as Rameau, but the difference is one of degree, rather than the difference in kind that stands between us and Rousseau as he believes himself to be. We do have characters which are steady to some degree and we do have some settled basis of motivation and belief on which to base practical decisions, but we should not imagine that these are fixed, and should also recognise that their possession is itself an attainment.

Williams goes on to explore what is happening when our minds become steadied; when we move from a position of uncertainty to the attainment of character or the establishment of a practical position. The resulting picture contains two main elements: those primitive contents of the mind which precede beliefs and desires; and what happens when they become beliefs and desires. We experience the contents of the mind that precede fully formed desires and beliefs whenever we embark on any instance of deliberation in which we are not entirely sure of our ground. Of course, this includes many, if not most, instances of deliberation, including those instances we might best describe as idle speculation or daydreaming, as well as those involving focussed, serious thought. Our experience is that, although such deliberation may include the consideration of recognisable and well understood desires and beliefs, we also find ourselves immersed in half-formed, unarticulated thoughts and wishes which are insufficiently defined to be regarded as beliefs and desires. These fragmentary and sometimes elusive elements do not comprise the sole raw material of deliberation, but they are sufficiently involved that

[54] *Truth and Truthfulness*, page 193.

any full picture of deliberation must take account of them. As Williams says, 'the process of arriving at a practical conclusion typically involves a shifting and indeterminate set of wishes, hopes and fears, in addition to the more clearly defined architecture of desire and belief.'[55] Williams sees the way in which these shifting and indeterminate contents of the mind become settled as one of *commitment*. That is, as these contents of the mind present themselves to us we find that we can more readily commit to some rather than others, and it is this commitment which establishes them as beliefs. Because Williams is concerned with the expression of the truth, he concentrates on the tendency of trustful conversation, in which we are often obliged to say what we believe, to produce such commitments: 'At a more basic level, we are all together in the social activity of mutually stabilising our declarations and moods and impulses into becoming such things as beliefs and relatively steady attitudes.'[56]

There are two further aspects of Williams' picture of steadying the mind which we should note, one of which he states explicitly, and the other of which is implied. Firstly, it is possible for steadying the mind to go wrong. Just because the primitive contents of the mind *can* become settled into beliefs and desires does not mean that they *should* become so settled. Williams distinguishes carefully between desires and wishes, characterising wishes as motivational states whose content it may or may not be possible to satisfy within a particular practical context, and further categorising *mere* wishes as those motivational states which we know cannot be satisfied within a particular practical context.[57] Steadying the mind goes wrong when mere wishes become established as desires or, worse still, as beliefs, resulting in deliberation which is inevitably futile. As Williams says, this, 'suggests that wishful thinking is not at all mysterious. It turns out, in fact, to be precisely well-named – it is thinking which is full of wishes, and, since all practical thinking is full of wishes, in the most general sense of the term in which wishes can occur on the route both to belief and to desire, there is no mystery about the fact that (to put it crudely) an agent may easily find himself committed to their content in the wrong mode.'[58] The second point to note about steadying the mind is that, although Williams does not say so explicitly, it is clear that he does not intend that it be understood as an activity which we undertake, either consciously or unconsciously. We do not set out to steady our minds; rather, we engage in deliberation, or attempt to express to somebody else what we want or what we believe, or get on with our practical lives in any of the innumerable ways which are available to us, and our minds become steadied as a by-product.

The concept of steadying the mind has consequences for Williams' argument in *Truth and Truthfulness* and for our discussion in three ways. Firstly, it represents a development in Williams' construction of an *imaginary genealogy*. This method, which is more frequently encountered in political philosophy, although it is also used by Hume in his account of the virtues, involves speculation about the character traits we could expect to emerge in any group of human beings living in a primitive state labelled the State of Nature. The State of Nature and the people living in it are entirely fictional, hence *imaginary* genealogy, but considering how they would develop and interact given what

[55] *Truth and Truthfulness*, page 198.
[56] *Truth and Truthfulness*, page 193.
[57] See *Truth and Truthfulness*, page 195-196.
[58] *Truth and Truthfulness*, page 197-198.

we know about actual human beings can tell us something about the basis of our own values and behaviour. The particular imaginary genealogy constructed by Williams in *Truth and Truthfulness* is not of great interest to us now, but we will find the method of use later in our discussion. The second consequence of the concept of steadying the mind concerns the development of character. As we have seen, the concept implies that our characters are developed as we become steadied through living our lives, and particularly through our engagement with others, and that consequently we should not conceive of ourselves, as Rousseau does, as possessing thoroughly settled characters which we reveal through our actions and utterances if we live authentically, but which persist even if we consistently betray them by living inauthentically. Rather, we construct our characters as well as revealing them through our actions and utterances.

The third consequence of the concept of steadying the mind is the most relevant to our discussion, and is that it modifies our understanding of what goes on in particular instances of deliberation. This is of obvious importance to our discussion of internal reasons, particularly when we consider how our desires become settled to the extent that they are capable of giving us reasons for action. Unfortunately this is not an area to which Williams gives a great deal of attention in his discussion of steadying the mind: all he does is to explore the distinction between desires and wishes[59] and to point out that desires may become settled in a fashion similar to beliefs.[60] Fortunately these thoughts, combined with the more general understanding of steadying the mind we can gain from the rest of Williams' discussion, as well as the thought we have already encountered in Williams' work on internal and external reasons that the contents of an agent's subjective motivational set, S, may be modified by deliberation, together give us enough to construct a picture of steadying the mind as it relates to motivations and the reasons which depend on those motivations.

We take four essential elements from Williams' general account of steadying the mind: the thought that the contents of our minds are neither entirely settled nor entirely transparent to us; the observation that when we deliberate we are often dealing with part-formed thoughts and feelings which are not sufficiently defined to be regarded as desires or beliefs; the distinction between these part-formed thoughts and feelings and firmer mental contents such as desires and beliefs not just through the degree to which they are understood and articulated, but also through the degree to which we are committed to them; and, finally, the way in which engagement with others acts as a catalyst for steadying the unsettled contents of the mind. When we consider these elements in the context of the discussion of internal and external reasons a picture emerges in which the subjective motivational set, S, is not only partly obscure to us but is, in its least well defined reaches, associated with mental contents which could not yet be properly considered as motivations, but which could potentially become motivations as the mind is steadied. These mental contents may become steadied into motivations when the agent is placed in a situation where it is imperative to decide what he or she really wants; to commit to desires in areas where commitment may never have been required before. This commitment may be required because of the need to engage practically with others but, importantly, it may be required even in instances of solitary practical deliberation where a lack of commitment would prevent the agent from reaching a practical conclusion.

---

[59] *Truth and Truthfulness*, page 195-198.
[60] *Truth and Truthfulness*, page 193.

An illustration may be helpful here. Imagine a woman who is successful in business, who also has strong political views and who, as a consequence, has from time to time considered the possibility of running for political office. However, to date this has remained an idle speculation; her thoughts regarding a potential political career have remained firmly in the category of wishes. Now, though, she has been approached by a local constituency organisation to ask whether she will stand as a candidate. Furthermore, this unsolicited approach is a sign that her candidacy is sufficiently valued that she could anticipate a political career as successful as her business career. She is naturally going to deliberate about what she should do. Some of this deliberation will be fully articulated as an explicit argument. Much of it, however, will be imaginative, and will involve the 'shifting and indeterminate set of wishes, hopes and fears' discussed by Williams. In a less extreme sense than that applicable to Rameau, the agent will determine how she defines herself and presents herself to others according to how she makes this decision; as a politician or as a business person. The situation demands a shift in the level of commitment made to previously vague wishes to pursue a political career: either they achieve the status of definite desires, with consequences for the agent's practical reasons; or they are relegated to the status of mere wishes which, for the time being at least, have extremely limited implications for the agent's reasons and actions. Furthermore, the process of deliberation may settle or modify other actual or potential elements of S which are more peripheral to the matter under consideration. In considering her options the agent may wonder whether she could tolerate the levels of sycophancy and dissembling which often seem linked to political success, and may find that, by imagining herself in situations where such behaviour is politically expedient, her attitude towards it changes and becomes more clearly defined: she may come to see it as a necessary part of achieving political goals; or she may see it as a compromise of her integrity that she could never accept. Whatever her decision, she may find her attitudes towards politicians and the political process become rather different and rather more definite than before.

So, it seems that the concept of steadying the mind can be applied to the development, definition and modification of the various elements of S through individual instances of deliberation as well as those areas which Williams explores more fully. We should note, however, a couple of places in which our brief discussion has a different emphasis from Williams' account. Firstly, we are primarily concerned with desire rather than belief, and this leads us to observe that steadying the mind seems to produce a finer grained distinction within what Williams calls the 'register of desire' than within what he calls the 'register of belief'.[61] When primitive contents of the mind becomes steadied into beliefs, it seems most natural to use this single term to label them. We may talk of strong beliefs or tentative beliefs or even passionate beliefs, but we are still talking about beliefs. By contrast, although in philosophy we may use some general term such as desires or motivations or even content of the subjective motivational set S to label mental contents with a predominantly motivational aspect, this is always an uneasy compromise; we are always aware that the term is standing in for a range of everyday terms, including desires and motivations, but also including wants, needs, dispositions, hopes, fears, inclinations, likes, loves and so on. This indicates to us that the process of steadying the mind does not necessarily stop when the contents of the mind which precede desires and beliefs have been settled into beliefs and the range of mental contents which we label desires.

[61] *Truth and Truthfulness*, page 197.

Steadying the mind may also involve shifts, whether radical or subtle, within the register of desire; wants may become needs that we are less capable of resisting, repeated instances of desire may become so common and familiar that they become further settled into more general dispositions and so on. This observation does not challenge or disrupt Williams' account of steadying the mind, but does remind us that the contents of S can remain shifting and indeterminate even after they have been established as contents of S.

The second difference in emphasis between our employment of the concept of steadying the mind and Williams' account is that Williams is primarily concerned with our engagement with others, while we are primarily concerned with individual instances of deliberation, and, while deliberation can be conducted with others, we think of it first as a solitary activity. This does not mean that we cannot use the concept of steadying the mind within our account, though. I believe that much single person deliberation resembles a dialogue conducted with oneself to find out what one really thinks or wants, and that this dialogue is analogous to engagement with others. We must be careful here: Williams specifically warns against the Platonic conception of the mind as an assembly of competing voices[62] and points out that such models soon run into trouble as we start to see the competing voices within the agent as potential personalities in their own right. But this model is not what I am proposing. Such an image is flawed because it supposes that we possess a determinate set of beliefs and desires which clamour for our attention. Our picture is of a much more tentative and exploratory process: we imagine our agent who, when confronted with the shifting set of considerations which arise in deliberation, asks him or herself whether he or she believes or feels them, not in the sense of discovering the truth of a pre-existing fact, but in the sense of catalysing the process of settling what is *going to be* believed or felt. Indeed, I think that this picture, far from challenging Williams' more socially grounded account, allows us to see how this account can be extended to instances of single person deliberation. Mature deliberating adults are engaged in a more or less constant dialogue with themselves which resembles their engagement with others, and which consequently demands that they settle their beliefs and desires in order to answer the practical questions by which they find themselves confronted.

Of course, although this may be a new way of articulating it, the picture we have drawn is not unfamiliar: it is the same as that in our account of internal reasons, in which deliberation may influence the contents of the starting motivational set. The only difference is that the starting point is rather more basic and the influence is rather more profound than we had originally imagined. So, Williams' account of steadying the mind appears entirely compatible with our account of internal reasons, and as it accounts for an essential aspect of the relationship between deliberation and motivations, it should be incorporated within our account. As we have said, there is nothing in this account which presupposes a fixed starting point, although it may be easiest to imagine one. We can allow that the starting point of deliberation is unsteady to a degree, although it will be settled in deliberation. The main consequence of introducing the concept of steadying the mind is that, as well as enriching our picture of motivations, deliberation and reasons, it increases the degree to which we understand reasons to be indeterminate. If the process of deliberation steadies the motivations on which reasons depend, and if this steadying is not itself fully determinate, then we cannot say ahead of the agent's deliberation exactly what

[62] *Truth and Truthfulness*, page 194-195.

reasons he or she has. Of course, we can predict many of the agent's reasons – we are sufficiently alike for that – but we cannot exhaust the agent's reasons. So, the picture implied by our modified account of internal reasons allows that, as well as picking our way through a landscape of desire and belief of which we are relatively sure, we are also often dealing with contents of the mind which are rather less well-defined, and which will only become so through deliberation.

Of course, we must acknowledge that, by introducing the concept of steadying the mind, we have once again modified our account of internal reasons in a way which seems to satisfy the individualistic intuition about reasons but which also seems, at least superficially, to confound the universalistic intuition. Not only are we maintaining that reasons are dependent on factors such as motivations which are peculiar to individuals, we are saying that these factors may not even be settled within those individuals prior to deliberation. The hope of making judgements of reasons which can be applied to agents even if they do not acknowledge those reasons may seem to fade in the face of this thought. However, in our earlier discussions of Williams and Hume we have noted that allowance for the obscurity of reasons creates room for at least one way of satisfying the universalistic intuition; by holding out the prospect of meaningful judgement and engagement concerning reasons, even with those people whose reasons seem at first sight to be at odds with ours. The idea that motivations are not settled entirely before deliberation further extends the possibility of engagement and persuasion. Agents may not just fail to apprehend reasons because their deliberation has gone wrong; their motivations may not even be settled to the extent that these reasons yet exist for them. The implications of this are perhaps clearest when we consider people who could be described as shallow, or as lacking in seriousness about certain aspects of the world (accusations which I suspect could legitimately be levelled at all of us in some respect). For example, imagine those exasperating people who claim not to be interested in politics because, 'politicians are all the same,' or because, 'they're all in it for themselves.' Sometimes such attitudes may well be thought out positions supported by stable underlying motivations, in which case we will probably want to find out exactly what those positions involve. More often, however, I think that we will find that the agent has *no* firmly settled motivations or beliefs in this area, and has never conducted the deliberation or developed the interest that would settle them; we are familiar with the thought that people who had never engaged with the political process may become surprisingly committed and radicalised by experience of conflict or crisis. This does not mean that it will be easy to satisfy the universalistic intuition in individual encounters with apathetic or shallow agents, but it does at least hold out the possibility. Of course, once more, we must acknowledge that this particular way of satisfying the universalistic intuition will not be enough for everybody.

1.5 Obstacles to Deliberation

In our discussions of Hume and Williams we observed that the recognition of obstacles to deliberation may satisfy the universalistic intuition about reasons as well as the individualistic intuition, because it implies that judgements of reasons may be legitimate even if the agent does not apprehend those reasons; it leaves the possibility of engagement open, even with an apparently intransigent interlocutor. However, we must also acknowledge that the implications of this thought for our talk of reasons are far from obvious or clear. Although Hume and Williams did not express this thought explicitly, they did recognise the ambiguity it produces. So, Hume allows that passions are unreasonable when based on the existence of objects which do not exist or on the mistaken belief that certain means will produce certain ends, while Williams formally includes the consequence of false beliefs within his account:

> '(ii) A member of *S*, *D*, will not give *A* a reason for ϕ-ing if either the existence of *D* is dependent on false belief, or *A*'s belief in the relevance of ϕ-ing to the satisfaction of *D* is false.'[63]

However, Hume goes on to say, with respect to passions based on false judgements that, 'even then, 'tis not the passion, properly speaking, which is unreasonable, but the judgment.'[64] Similarly, Williams says, with respect to the man who wants a gin and tonic and is unaware that his gin bottle contains petrol, 'On the one hand, it is just very odd to say that he has a reason to drink this stuff, and natural to say that he has no reason to drink it, although he thinks that he has. On the other hand, if he does drink it, we not only have an explanation of his doing so (a reason why he did it), but we have an explanation which is of the reason-for-action form,'[65] and later, after introducing element (ii) of his scheme which we saw earlier: 'It will, all the same, be true that if he does ϕ in these circumstances, there was a reason why he ϕ-ed, but also that displays him as, relative to his false belief, acting rationally.'[66] It appears that there is something about deliberative obstacles such as ignorance which not only makes them loci of tension between the individualistic and universalistic intuitions about reasons, but also confounds our intuitions about what to say. Hume does not pursue this question further, and Williams attempts to close it off by saying that, with regard to the question of whether we should say that an ignorant agent has a reason to act as if he or she were not ignorant, or whether we should say that an ignorant agent would have such a reason if he or she was not ignorant, 'I shall not pursue the question of the conditions for saying one thing or the other, but it must be closely connected with the question of when the ignorance forms part of the explanation for what *A* actually does.'[67] I think that we can and must be a little bolder than Hume or Williams; the question of whether to say one thing or the other, or rather the question of whether this is the right question at all, is one that we cannot afford to leave unexplored. In order to pursue it, and to establish whether our suspicion that

---

[63] 'Internal and external reasons', page 103.
[64] *Treatise*, Book II, Part III, Section III, page 463.
[65] 'Internal and external reasons', page 102.
[66] 'Internal and external reasons', page 103.
[67] 'Internal and external reasons', page 103.

deliberative obstacles help rather than hinder the satisfaction of the universalistic intuition about reasons is correct, we must examine the concept of the deliberative obstacle a little more closely.

We shall start by considering the types of deliberative obstacles that exist in addition to straightforward ignorance. These obstacles fall into three categories: those associated with deliberative habits; those associated with deliberative capacities and capabilities; and those associated with irrationality. Each of these sets of obstacles causes us problems; even the last, as we must show that however we account for the other forms of deliberative obstacles, we still allow that sometimes we are irrational. That is, if we find grounds for excusing a failure to follow or avoid reasons due to the presence of deliberative obstacles, those grounds must not excuse the agent who is simply acting irrationally.

We can gain an understanding of the role played by habits in deliberation by imagining the nearly infinite variety of ways in which deliberation could go from any point where an agent has to make a practical decision. In the example of the twin brothers we used earlier on to illustrate the meaning of indeterminacy we considered only two possible courses of action: capitulation or resistance. However, there were many other actions available, some likely to occur to someone in a similar situation, but others unlikely to occur to anybody. So, the brothers could have considered fleeing the country altogether, either to start a new life, or to join forces with the foreign enemies of the invader. The more unlikely courses of action include betraying their home country and joining the enemy, killing themselves to avoid their dilemma, throwing away their lives in a suicide mission, abandoning human company to live as a hermit, and, indeed, anything else physically possible in their circumstances. Of course, in our example they didn't think like this, and in our everyday lives we don't think like this. Habits guide us down tested deliberative paths. The idea that we don't consider every deliberative option available to us is illustrated by the apparent differences between the way in which expert human chess players make their decisions, and the way in which chess computers work.[68] While chess computers grind through every possibility, evaluating all potential outcomes, humans apparently only consider those sequences of moves which they know are worthwhile. Of course, this raises the possibility, however unlikely, that the process of considering every option will discover an unexpected move superior to that which can be produced by intuition guided by habit.

The most influential habits are those which are so ingrained in us that their operation is invisible. This includes the performance of trivial tasks such as making a cup of tea or tying a shoelace, in which we never stop to consider whether there might be a better way, but also includes more pervasive deliberative habits, such as our tendency to worry about our everyday problems without wondering whether we could so reorder our lives as to make these problems irrelevant. When beset with financial, family or work crises, the thought that they could all be escaped by adopting a simpler mode of life may occasionally cross our minds, but not to the extent that it could be considered a potential course of action that we actually deliberate about. And for most of us, most of the time, these constraints on our deliberation are entirely appropriate; our modes of life come with problems which can usually only be addressed by deliberation which takes place within

[68] Or at least worked. A better understanding of way in which humans play chess influences the development of computers.

the boundaries of those modes of life. However, while they may be the most prevalent and powerful, invisible deliberative habits are the most difficult to analyse, precisely because they are invisible. If we do try to analyse them we may find ourselves guilty of invention rather than analysis. Consequently, we will look for more explicit examples of deliberation guided by habit. In particular, we will look at three types of habit: individual, institutional and cultural.

Individual habits are those in which an individual has established a deliberative route to a particular conclusion in response to a particular problem, and reuses that route whenever that problem or anything related to it arises, possibly to the extent that the problem is no longer perceived as a problem. For example, imagine a man who has moved to a new town to take up a job. One of the mundane problems which faces him is the best way to get to work every morning. When the town is new to him he may explicitly deliberate about this problem, consulting his colleagues, looking at maps, and trying out different routes. Eventually, though, he will settle into a routine, and the question of how to get to work no longer presents itself; he simply follows his habitual route. What makes this a deliberative habit as well as just a habit, is that when deliberation does take place it will do so in the context of the established solution. If he unexpectedly encounters roadworks one day he is likely to think of it as an inconvenience, even if he is forced to take a route which actually turns out to be slightly better. The irritation of disruption to his habit obscures his perception of reasons.

Institutional habits are those in which an institution of which the agent is a member has reached a particular conclusion about how to tackle a particular problem, and embedded it in the way it works. Such habits may determine specific procedures, but may also include the ways in which the institution identifies problems, debates solutions and makes decisions. The persistence of institutional deliberative habits is illustrated by the story 'Chromium' in Primo Levi's book *The Periodic Table*. In this autobiographical story Levi tells of the time when, while working as an industrial chemist in a paint factory, he was presented with the problem of why some batches of paint set like jelly; a phenomenon known as 'livering.' He discovered that the problem was caused by errors in the quality control procedures used to check the raw ingredients, and returned the livered paint to its liquid form by adding ammonium chloride to the mixture. Years later he heard from a friend that the technicians at the factory were still following his instructions and adding ammonium chloride to the paint, even though the quality control procedures had been corrected and, far from being required any more, ammonium chloride was even harmful to the paint's purpose as an anti-rust coating. The technicians following Levi's instructions didn't know the original reason for them; all they knew was that that was the way it had always been done. Although the technicians had reasons for acting as they did, they were prevented from deliberating to other reasons by the habit of following Levi's instructions.

Cultural deliberative habits are those which are embedded in the culture in which the agent lives, and which are transmitted by cultural mechanisms such as education and socialisation. Cultural deliberative habits are distinguished from other cultural values and conventions because they concern ways of thinking about problems and reaching practical solutions. Such habits may be highly explicit and formalised, such as the laws and constitution of a republic, or they may be implicit and informal, such as the unspoken rules governing acceptable behaviour in certain social contexts. The important point is

that they are typically not acquired by the agent consciously through rational means, yet constrain the practical options which an agent will consider.

So, whichever of these categories a deliberative habit belongs to, it has a similar effect on our practical reasoning. By keeping our deliberation running in a certain groove, it allows us to apprehend satisfactory reasons without having to consider every available option, but also prevents us from apprehending those reasons which would require a more unorthodox deliberative leap. However, while we can see how such habits might get in the way of sound deliberation, it may seem unlikely that are of genuine significance for our account of internal reasons. After all, such habits do not seem to present the same insurmountable obstacle to deliberation as that presented by ignorance. There is no amount of mere reasoning that will tell me that a gin bottle contains petrol; I have to open the bottle to find that out. By contrast, we can see how deliberative habits could be broken by reasoning alone, especially if we realise that an agent does not have to shed a habit entirely to break through it from time to time to reach a reason that would otherwise be obscured. However, we must not underestimate the power and ubiquity of deliberative habits. It is not only in such dramatic situations as that of the twin brothers in our example that we constrain our practical options to those which seem of particular relevance and importance. At every point of practical decision in our lives, we could conceivably consider an infinite range of options, but to do so would result in practical paralysis. This is not just a question of whether we have time to consider every option; it is rather that we are simply not capable of deliberating like that. Like the chess player, we see certain patterns in our practical options which are worth pursuing, and we do not even consider the others. We are operating in a rather more complex domain then the chess player, and the likelihood that we ignore options which could be better than those which we do consider is rather higher. So, deliberative habits are not merely exceptional considerations which occasionally interfere with the operation of our reasoning; they make it possible for us to reason about practical matters and to reach conclusions at all. Nevertheless, they can present genuine obstacles which are as hard for agents to overcome as ignorance.

However, if it is still difficult to accept that deliberative habits present such obstacles, it should be easier to accept that deliberative capacities and capabilities – or a lack of them – do present such obstacles. The presence of a habit implies that an agent deliberates in a particular way, but could conceivably deliberate in a different way. The lack of a capacity or a capability implies that the agent could not deliberate that way at all. However, although they have similar implications for reasons, capacities and capabilities are not quite the same as each other, and we must distinguish between them carefully. By capacities we mean aptitudes for deliberation which are innate, although they may be revealed and developed by experience. Examples may be as specific as a knack for solving crossword puzzles, or as general as the possession of an active imagination. When we think of capacities in the context of deliberative obstacles, we may most readily think of those capacities which most of us possess but which some people conspicuously lack, such as people whose brains have been damaged in some way. However, we must also recognise that some people have exceptional capacities which most of us lack. An extreme example of deliberative capacities which are beyond most of us can be found in James Gleick's biography of the physicist Richard Feynman, *Genius*. Gleick quotes an interview in which Feynman tries to explain how he imagines the

behaviour of the world at the sub-atomic level of quantum physics; behaviour which is nothing like that of the world at the level we are used to, and which even those people who can describe it in mathematical terms find unimaginable in physical terms. Feynman says, 'What I am really trying to do is bring birth to clarity, which is really a half-assedly thought-out pictorial semi-vision thing. I would see the jiggle-jiggle-jiggle or the wiggle of the path. Even now when I talk about the influence functional, I see the coupling and I take this turn – like as if there was a big bag of stuff – and try to collect it away and to push it. It's all visual. It's hard to explain.'[69] There is evidently some form of deliberation going on here, but it is a form which seems as hard for Feynman to describe as it is for the rest of us to follow. But we do not have to look for such exotic examples of deliberation to find that which is beyond us. The feeling that answers to practical questions are just out of our reach, but that we could reach them if we were only better deliberators is a common one in many areas of our lives.

Capabilities are rather different because they may be acquired or taught. There is an obvious relationship between capabilities and capacities: capacities are often necessary for the achievement of capabilities. With practice I can get better at crossword puzzles, but without the capacity to decode anagrams or spot tricks of language, I will never have exactly the same capability as someone who possesses this capacity. This relationship is not entirely straightforward, however. Deliberative capabilities may include tricks and techniques which compensate for a lack of capacity. The capability of mental arithmetic for example, may benefit from the capacity to carry out lightning calculations, but for most of us is just as likely to rest on multiplication tables learnt by rote. The important characteristic of deliberative capabilities is that, once acquired, they give us new ways of thinking about the world and our practical problems. Consider the process of learning to drive, for example. This is a largely mechanical skill, but in learning to drive we start to see the environment around us in different ways: a simple practical question such as, 'How do I get there?' has new answers and new ways of going about divining those answers.

The existence of deliberative capacities and capabilities, and the possibility that any individual agent may lack them, presents us with a similar situation to that presented the existence of deliberative habits. We may be able to see a sound deliberative route between an agent's motivations and a particular practical conclusion, but because of the lack of a capacity or a capability, it is not possible for the agent to reach that conclusion, and no way in which action on the basis of that conclusion could be explained. We are left once again with conflicting intuitions: an intuition that a reason exists regardless of the impossibility of the agent deliberating to it; and an intuition that to ascribe this reason to the agent or to pass judgements of irrationality would be unjustified, just because of this impossibility.

The situation is further complicated by the possibility that the capacity lacked by the agent is the capacity for rationality, and that in that case judgements of irrationality would be justified. Irrationality is troublesome for our account, because deliberation which is obstructed by irrationality seems to possess many of the same characteristics as deliberation blocked by the other obstacles we have considered. However, while we want to allow that the existence of other types of obstacle complicates the question of whether particular reasons exist or not, we do not want to extend this same allowance to the case

[69] *Genius: Richard Feynman and Modern Physics*, page 244.

of irrationality: we do not want the agent's irrationality to determine the existence of his or her reasons. In order to account for irrationality adequately it is important to be clear about what we are talking about. Part of the reason for our entire discussion is that the term is often used as a non-specific pejorative term: it may be part of a bluff, an insult or another judgement of disapproval, rather than a judgement about whether the agent is acting in accordance with his or her reasons. I believe that, within the context of our account, we can most usefully define irrationality in terms of deliberation, and in particular as the pursuit of deliberative paths in a fashion which cannot itself be justified by reference to reason. This may sound somewhat circular, but it captures the thought that deliberation is an action, albeit a predominantly mental one, and is subject to the same rational appraisal as other actions. In keeping with our account, and our understanding that reasons and deliberation are to a large degree indeterminate, it would not be possible to exhaustively describe or systematically circumscribe the appraisals of deliberation which constitute judgements of irrationality. However, we can identify those common circumstances in which we recognise irrationality.

Considered in this fashion we can identify three common categories of irrationality: simple error; disordered reasoning; and wilful disregard. Simple error occurs when we make a mistake in our reasoning, and most closely resembles the obstacles presented by lack of capacity or capability. Obviously, a lack of capacity or capability may also be instrumental in producing error. However, we must also acknowledge that whatever our capacities or capabilities, sometimes we just go wrong. For example, consider a man who has an unwelcome announcement to make to his family: his company is sending him abroad for the year. While considering how to break this news the man dwells on the positive aspects of the situation, until he becomes carried away. He decides that the best thing to do would be to present the news to his wife and children in a light-hearted, upbeat manner. But he has miscalculated terribly: his family are not only upset by the news, but find his flippant manner positively hurtful. He could and should have realised the consequences of his decision, but failed to think about it properly. He acted contrary to his reasons because he made an error in his deliberation.

If simple error seems on the border of irrationality, then disordered reasoning lies well within its territory; indeed, it could be regarded as the paradigm case of irrationality. By disordered reasoning we mean that the agent is prone to drawing connections and conclusions which have no rational basis. Such disorder may be caused in many ways, from mental illness to momentary bursts of emotion.[70] As an example, we may imagine a man in a stressful situation: he has just emerged from an interview with his boss in which he has been fired on what he believes are thoroughly unjustified grounds, and must now decide how to deal with the inevitably undignified task of gathering his possessions and leaving the building. If he was mentally ill, his reasoning might be so disordered that he decided on an entirely inappropriate and ineffective course of action, such as marching to the police station and demanding the arrest of his boss. Even such disordered reasoning is not entirely inexplicable; we can see how someone could get from notions of being treated unjustly to the thought that the institutional agents of justice should avenge that

[70] Of course, I am not saying that bursts of emotion are irrational in themselves, or that they necessarily give rise to irrational action. Rather, I am saying that they are capable of so gripping the agent's mind that sound deliberation is not possible. This is captured in such common sayings as, 'I was so angry that I couldn't think straight,' or simply, 'I saw red.'

treatment. But it can hardly be called rational. As it happens, he is not mentally ill, but is rather extremely upset and angry. As a consequence, he sees everyone associated with the company as implicated in his ill-treatment, and consequently not only rebuffs any offers of help, but is rather insulting to those who offer help. Later, when he has calmed down, he regrets his actions, and sees that, while he had every reason to be angry, he had no reason to treat his colleagues so badly.

We tend to feel that, to varying degrees, people carry little blame for disordered reasoning. We make allowances not just for those people who are mentally ill, but also those who are momentarily overwhelmed. In the situation we have just considered, we would expect that the man's colleagues would not only readily forgive his behaviour, but would find it, if not rational, entirely understandable. By contrast, we feel that people who exhibit the final species of irrationality we will consider – wilful disregard – particularly culpable. Wilful disregard occurs when someone deliberately ignores a deliberative path or conclusion, even though he or she knows it to be sound. For example, imagine a woman who is a professional athlete, and who has injured herself the night before a major event. She knows that if she competes she will not only lose, but will do herself permanent damage. However, she cannot bear to believe that her ambition will be thwarted and that her preparation will be wasted, so she competes anyway. We must be clear about her intent here: she is not competing as a futile or defiant gesture, but is competing as if she can win even though she knows that she can't. She is wilfully disregarding her reasons and any other consequences of her deliberation.

So, now that we have a clearer idea of the main varieties of irrationality, we can return to the problem from which we started. Our primary argument against the existence of external reasons is that there is no way in which the agent could act for such reasons, and we maintain that this has normative as well as explanatory implications. Deliberative obstacles prevent the agent from following certain deliberative routes, and similarly negate the possibility of acting for the reasons which lie at the end of those deliberative routes. For some of these obstacles – such as ignorance, habit and a lack of capacity or capability – our informal intuitions about the existence of reasons conflict: we want to say that reasons exist, even when they lie on the other side of deliberative obstacles; yet we also want to allow that in some sense the agent does not have these reasons, and that if an agent acts as if these reasons do not exist, he or she is nevertheless acting rationally. The vocabulary available to our account of internal reasons so far does not allow us to reflect this conflict: if the deliberative obstacles were constituted by possession of the appropriate type of false belief, the implication of both Hume's and Williams' arguments is that the agent nevertheless possesses the reason, and that we should judge his or her actions accordingly. For other obstacles – such as those associated with irrationality – our intuitions are clearer and in accordance with our account so far: we want to say that the agent has only those reasons which are not dependent on irrational beliefs or deliberation, and that the agent has those reasons regardless of whether irrationality prevents their perception. Our problem, then, is to extend our account to allow for the influence of deliberative obstacles on the existence of reasons in a way which satisfies our conflicted intuitions, but neither provides room for reasons to be based on irrationality, nor requires us to believe in external reasons.

Let us return to Williams' example of the gin drinker to see if he helps us to find our way through this problem. Remember that the man in this example wants a gin and

tonic, and believes that the bottle in front of him contains gin, whereas it actually contains petrol. We want to say that he has a reason to avoid drinking what is in the bottle, because it would obviously harm him, and if he has the normal human motivations concerning self-preservation, he has plenty of reasons to avoid harm. At the same time, if he cannot apprehend this reason due to ignorance (he doesn't know what is in the bottle), deliberative habit (he normally trusts the labels on bottles) , or a lack of capability (he has never learned to speculate suspiciously about the motives of people who offer him drinks) then we are also inclined to say that he has a reason to drink what is in the bottle. It is only when the obstacle is posed by irrationality that we are not inclined to do this. So, if the agent had committed a deliberative error (the smell of the petrol aroused his suspicions, but he reasoned that any bottle which is labelled 'gin' must necessarily contain gin), was suffering from disordered reasoning (he is an alcoholic whose cravings are so strong that he is prevented from contemplating any possibility that he is not going to get a drink soon), or was wilfully disregarding the conclusions of reason (the smell and the appearance of the liquid are unmistakably petrol, but he deliberately ignores this sensory evidence because he so strongly wants a drink), we would not be tempted to say that he had a reason to drink what was in the bottle: we would just say that he was irrational.

This example helps us to see that, in the cases of deliberative obstacles other than irrationality, although the conclusions reached by the agent about his reasons contradict those he would reach if he was better informed, they can nevertheless be reached through sound deliberation, exercised with all of the materials available to the agent. The obstacles represented by ignorance, habit and a lack of capacity or capability stand between the agent and his reasons precisely because they create deliberative routes which, although they do not lead to the best conclusions about action, are sound. We can say that such deliberative routes are sound because, although they lead to less than perfect conclusions, they depend on deliberation which, *as deliberation*, has nothing wrong with it. As we have noted, it is not possible for the deliberating agent to pursue every deliberative path which is conceivably available, so the paths chosen are necessarily constrained by knowledge, habits, capacities and capabilities. Indeed, in many cases, deliberation which surmounted such obstacles would have to do so by being unsound. By contrast, the deliberative obstacle of irrationality stands as an obstacle because it leads the agent down deliberative paths which are by definition unsound. And our account already rules out reasons which are produced through unsound deliberation.

So, deliberative obstacles both obscure reasons that could be reached if the obstacles did not exist, and can lead us to settle for reasons which we would not accept without the obstacles. As long as we retain the concept of the sound deliberative route as a pre-requisite for reasons in mind, then we can account for our troublesome intuitions concerning this particular case without endorsing either irrationality or external reasons. The intuition that the gin drinker had a reason to drink what is in the bottle concerns the reasons he is capable of reaching as an ordinary deliberator, subject as we all are to deliberative obstacles. The intuition that he had a reason not to drink what is in the bottle concerns the reasons he would be capable of reaching if he was not subject to such obstacles. We can satisfy these intuitions within our account and retain the central claim that the existence of reasons is dependent on the possibility of sound deliberation from an agent's starting set of motivations: irrationality does not satisfy the condition of sound

deliberation, and external reasons do not satisfy the condition of dependence on motivations.

Furthermore, this deeper understanding of deliberative obstacles, and the role played by sound deliberation in distinguishing between obscurity and irrationality confirms our suspicion that allowance for deliberative obstacles helps satisfy our universalistic intuition about reasons without confounding our individualistic intuition. It does so in two ways. Firstly, it allows for a wide variety of ways in which reasons can be possessed by an agent yet be obscure to that agent. The possession of such reasons may still be entirely dependent on the agent's motivations, character and relationships as the individualistic intuition wants it to be, yet the existence of deliberative obstacles may mean that such reasons are only apparent from the perspective of an external observer. Someone who is apparently attempting to convince a recalcitrant agent that he or she possesses a reason may well be bluffing, but if the persuasion consists, as it often does, of attempts to dispel ignorance, or to establish a deliberative capability, or to compensate for a lack of deliberative capacity, or to break through ingrained deliberative habits which were not previously apparent then it is not bluff. The second way in which our understanding of deliberative obstacles is capable of satisfying the universalistic intuition lies in our continuing insistence on the existence of a sound deliberative route as a pre-requisite for the existence of reasons. Even though we understand that what counts as sound deliberation is not fully determinate, it means that the part of our universalistic intuition which expects that it must be possible for reasoning to go wrong in ways which are identifiable as such independently of the deliberating agent can be satisfied. Furthermore, the existence of this independent standard does not contradict our individualistic intuition; wherever sound deliberative routes end up, we insist that they start with the motivations of the agent.

We are still left with the problem of what to say, though; we must decide how to talk about these different categories of reasons, those which we can actually reach through deliberation and those which we could reach if only we were not beset by obstacles. In everyday speech, of course, we just call them all reasons, and insofar as our meaning is made explicit, it is given by context. Because we have not yet identified more precise equivalents in everyday language, I shall draw a technical distinction between two types of reasons:

> *strongly internal reasons*: those reasons which *can be* apprehended through sound deliberation by the agent, and;
>
> *weakly internal reasons*: those reasons which *could be* apprehended by the agent if not for the existence of deliberative obstacles.

The most important implication of this categorisation is that judgements of reasons are even more contingent than we originally supposed. Our account has always implied that reasons were contingent on motivations and circumstances, as well as the different ways in which deliberation happens to go. We are now also saying that the reasons for which an agent can act are contingent on all these elements and also on the deliberative obstacles which the agent happens to face.

So, despite the initial awkwardness of our intuitions regarding situations in which

deliberation is blocked by obstacles, we have found that we can account for them by sticking closely to the core of our account. The implications of the existence of deliberative obstacles has led us to draw a distinction which, despite its unlovely terminology, supports the case for internal reasons, reconciles it with our intuitions, and reinforces the concepts of contingency and indeterminacy. However, we must recognise that this account is unlikely to satisfy everybody who attempts to understand practical reasons, and that consequently there are serious challenges which we must address.

1.6 Challenges

Let us review the picture of the agent, motivations, deliberation and reasons which we now have. We started with a fairly straightforward picture, in which we had a starting point, a set of possible routes from that starting point, and reasons at the end of those routes. Although we have not strayed far from Williams' account of internal reasons which gave us this picture, nor from its Humean origins, we have developed it in two important ways. Firstly, we have now complicated that picture by exploring and drawing together the implications of thoughts already present in Williams' work: that the routes available from a starting point will vary from agent to agent depending on the deliberative obstacles they confront; and that the starting point for deliberation itself is often not settled. We still maintain that reasons are dependent on motivations and are subject to judgements of what counts as sound deliberation. However, we now also allow that reasons are even more contingent and less fully determinate than we originally supposed. They are dependent on the agent's motivations, the agent's circumstances, the degree to which the contents of the agent's mind are settled and the processes by which they become settled, the deliberative obstacles which the agent happens to face, and the ways in which deliberation happens to go. Once again, this does not mean that reasons are entirely free; the criterion of sound deliberation continues to impose its constraints, and many of the new factors we have considered constrain reasons as well as producing difference. However, the range of possibilities within these constraints is much wider. Secondly, we have introduced the two differing intuitions about reasons, and have shown how, even though the internal reasons account naturally seems more friendly to our individualistic intuition about reasons, it is also capable of satisfying our universalistic intuition to some degree. We showed this by extending Hume's and Williams' somewhat abbreviated exploration of deliberative obstacles and argued that acknowledging such obstacles, despite their tendency to introduce further differences in the reasons of agents, allows for the possibility that agents have the reasons that we want to ascribe to them from an external standpoint, even though those agents do not realise that those reasons exist. So, in attempting to persuade others of their reasons we may be doing more than just bluffing, even if the subject of our persuasion is not convinced.

Of course, this is only one way of satisfying the universalistic intuition about reasons, and we must admit that it is not likely to satisfy everybody. However much allowance our account makes for the agent who shares our reasons but does not realise it, it also allows that sometimes an agent apparently doesn't share our reasons because he or she actually doesn't share our reasons; he or she just doesn't have the requisite set of starting motivations, and no amount of persuasion or heavy lifting of deliberative obstacles is going to reveal an unsuspected deliberative route. Our account allows that sensible knaves sometimes exist, and this is something that some theorists simply cannot accept. The only way that the universalistic intuition can be satisfied for such theorists is by demonstration of the existence of truly universal reasons which no agent can escape merely through the happenstance of his or her existing motivations. For such theorists concepts such as indeterminacy, contingency and deliberative obstacles comprise just so much fog that not only gets in the way of the apprehension of the universal reasons an agent is subject to, but also gets in the way of understanding the true nature of reasons. Such theorists are, of course, external reasons theorists.

There are many arguments for external reasons and we do not have time to pursue all of them here. Rather, I wish to concentrate on a particular style of argument which I believe presents the most serious challenge to our account, partly because of its dominance in ethics, and partly because it exposes a vulnerability in our account. We have insisted that although reasons may be contingent and indeterminate, they are not arbitrary; they are constrained by common standards of reason such as the principles of logic and by what counts as a sound deliberative route, even if this is itself indeterminate. Even our emphasis on deliberative obstacles implies the existence of common standards of reason; the categorisation of obstacles as obstacles implies that there are some criteria that make them so. The challenges I wish to consider are those which take such thoughts further, and include arguments that because standards of reason are independent of our contingent motivations, they are capable of determining reasons which apply to us all regardless of our motivations. In other words, the theories which present the greatest challenge to our internal reasons account are those which find substantive reasons within the conditions of rationality. Needless to say, these theories have a rather different relationship to the individualistic and universalistic intuitions about reasons than our own internal reasons account. They take the universalistic intuition to have priority; indeed, they take it to be nothing more than an informal acknowledgement of the true nature of reasons. By contrast, they see the individualistic intuition as nothing more than a hope; a wish that we could come to internalise the universal reasons which apply to all of us to the extent that they become our intimate possessions. To these theorists, of course, it is regrettable if this wish is not satisfied, but it does nothing to free agents from their obligations to abide by universal reasons.

There are plenty of variations within this style of argument and we could not possibly consider them all here, just as we could not possibly consider all other varieties of external reasons theory. Instead we will limit ourselves to representative arguments in three areas, each of which bears directly on a central element of our account of internal reasons. So, we will consider: arguments which derive reasons from the agent's *identity* and his or her rational nature; arguments which derive reasons from the *principles* supposedly expressed in deliberation and action; and arguments which derive reasons from the rational ordering of *motivations*. Our initial aim in these discussions will be to identify and respond to the challenges presented in each area, and to determine whether our account of internal reasons survives the challenges. However, assuming that we are successful in this aim we will also pursue a further objective. Despite our opposition to their arguments, we may suppose that the external reasons theorists we will consider are by and large rational agents, and are, in making their arguments, acting for reasons derived from their motivations by deliberation. Their pursuit of these arguments indicates that we need to do rather more to satisfy the universalistic intuition than to consider the implications of deliberative obstacles. Once we have dealt with the challenge posed by their theories, we will therefore also enquire what motivations might drive them to develop such theories, whether they might be shared in some form by people who do not express them through philosophy, and whether an understanding of them and the behaviour they produce might allow us to show how our account of internal reasons can more fully satisfy the universalistic intuition.

2. The Challenge from Reason

2.1 Identity

Arguments which attempt to derive external reasons from identity suppose that as human beings, or even just as rational beings, we possess inescapable aspects of our identities which are capable of providing us with reasons which are independent of our actual motivations. These reasons may either be provided as a direct consequence of possessing such an identity, or through the supposedly catastrophic consequences of denying that we possess such an identity. We shall consider two examples of this type of argument, one from Thomas Nagel and one from Christine Korsgaard. Before we consider these arguments, however, we should note two points about the way in which we shall proceed. Firstly, although both of these writers specifically criticise the claim that reasons can be derived from motivations, we shall not consider these criticisms here, even though I believe that we could provide at least an adequate response to them. The challenge constituted by these theories does not come from their negative arguments about why desires are incapable of providing reasons, but from the positive arguments which attempt to locate an alternative source of reasons. Secondly, I will not attempt to provide a point by point response to Nagel's and Korsgaard's arguments as we explore them, but will rather attempt to respond to their essential elements once we have laid the arguments out. In this fashion I hope to provide a general response to arguments which attempt to derive external reasons from identity, rather than responses which only deal with the particular arguments which we have chosen.

2.1.1 Nagel's Argument from Identity

Thomas Nagel has developed his argument that objective reasons can be derived from human identity over several years in various books and papers. His position and its development are best expressed by the books *The Possibility of Altruism* and *The View from Nowhere* and the differences between them.[71] In these books it is clear that Nagel is implacably hostile to the suggestion that reasons are derived from desires, partly because he believes that desires cannot produce the particular varieties of reasons he wishes to argue for, and partly because he believes that desires are generally inadequate as sources of reasons. However, although Nagel is hostile to the idea that reasons are dependent on desires, he does not deny our experience that reasoned action is accompanied by desire: 'The fact that the presence of a desire is a logically necessary condition (because it is a logical consequence) of a reason's motivating, does not entail that it is a necessary condition of the *presence* of the reasons; and if it is motivated by that reason it *cannot* be among the reason's conditions.'[72] In other words, he denies that desires precede reasons and that reasons then produce action, but allows that, although reasons precede desires, desires always accompany action. We should be aware, though, that Nagel sets out with a preconception about what constitutes a reason. He says that, 'Every reason can be formulated as a predicate. If the predicate applies to some act, event or circumstance

[71] But also see: the essay 'Subjective and Objective' in the collection *Mortal Questions*; *Equality and Partiality*, especially Chapter 2, 'Two Standpoints'; and *The Last Word*, especially Chapter 6, 'Ethics'.
[72] *The Possibility of Altruism*, page 30.

(possible or actual), then there is a reason for that act, event or circumstance to occur,'[73] and that, 'It is assumed, then, that reasons are universal – i.e. in some sense the same for all persons – and that they transmit their influence to actions suitably related to the ends to which they apply.'[74] This preconception that reasons are necessarily universal puts his theory at odds with an account such as ours, in which reasons are contingent on each agent's psychology and deliberative circumstances. As we shall see, it is also a major flaw in his theory, and one that will be repeated in the other external reasons theories we will consider.

Nagel's purpose in *The Possibility of Altruism* is to show not only that we are capable of being altruistic, but also that we all have reasons to be altruistic, no matter how we may happen to be motivationally constituted. Although our interest is not in the specific ethical point which Nagel is trying to make about altruism, we may note in passing that his ambition is a particularly strong expression of the universalistic intuition. Nagel evidently feels that we should all be more altruistic than we are, but he is not content that this simply stands as a moral judgement: through his theory he wants to show that we are all rationally obliged to be altruistic and that we are offending against our reasons if we fail in this respect. However, for the purposes of our current discussion the route which Nagel takes to his argument for altruism, which starts with the argument for prudence, is at least as important as the argument for altruism itself, because we are more interested in arguments for the existence of external reasons than in arguments about specific ethical points. Indeed, Nagel's argument for prudence may be of greater interest to us because, as we shall see, it is more plausible than his argument for altruism, and consequently presents the greatest challenge to our account.

The term prudence carries a variety of connotations in everyday language, mostly denoting caution and preparedness. The usual philosophical usage of the term is rather more precise, typically meaning an agent's concern for his or her own well-being. Nagel uses the term even more narrowly, to mean an agent's concern for his or her *future* well-being. So, in looking for prudential reasons, Nagel is not just looking for the reasons an agent has to get out of the way of immediate harm, but for the reasons which agents have to avoid getting into harmful situations at all. The prudential agent in Nagel's terms is not the man who buys a packet of cigarettes to satisfy his cravings, but the man who defies his cravings and gives up smoking altogether. And, on the face of it, this is an understanding of prudence which we can accept and sympathise with. Nagel claims that we all have prudential reasons. This is another claim which is extremely easy for us to accept, at least superficially, whether we are internal reasons theorists or external reasons theorists. Concern for the future is virtually ubiquitous in some form among reasoning adults, and those people who exhibit absolutely no concern for the future are often paradigm examples of irrationality. Even the person who commits suicide to avoid pain or degradation expresses a concern for the future; it is the man who risks his life for nothing who seems to lack such concern. However, while this ubiquity lends Nagel's argument credence, his insistence on our possession of prudential reasons is not simply based on our concerns for the future: if it was, then he would be accepting the existence of reasons based purely on motivations, which isn't what he is after at all. He must find something which gives us prudential reasons whether we care about the future or not. He

[73] *The Possibility of Altruism*, page 47.
[74] *The Possibility of Altruism*, page 90.

finds it in two rather more fundamental claims about the nature of practical reasons and the consequences of necessary aspects of human identity.

The first claim is an extension of Nagel's understanding of reasons as universal, and can be summarised as the claim that when we have a reason to act in the interests of an entity, we have a reason to act in the interests of relevantly similar entities. Put even more basically, if one thing acts as a source of reasons, then other things which are identical in relevant respects must also act as sources of reasons. If the interests of an entity are sources of reasons, then the boundaries and differences between entities must be shown to be relevant considerations if the interests of all entities are not also to be considered as sources of reasons. In the case of prudence the relevantly similar entities are the agent's future selves. So, Nagel is not going beyond self-interest here, but is maintaining that if the agent's self-interest provides reasons to act in the interests of his or her current self, then it provides similar reasons to act in the interests of his or her future selves. This is formalised further by distinguishing between 'tensed' and 'tenseless' reasons.[75] These labels are not meant to imply that there are distinct categories of reasons, but that reasons can have different relationships to points in time. A tenseless reason applies at any time, whereas a tensed reason only applies at a particular point in time. So, assuming that I want to avoid malaria, I have a tenseless reason to start taking anti-malaria tablets a week before a visit to India. This reason is tenseless because it applies whether I am going to India next week, next year or even if I have just come back. If I am planning a trip to India next week, however, then I have a tensed reason to start taking my anti-malaria tablets now. Nagel's articulation of the relationship between tensed reasons and tenseless reasons shows the general structure of his argument: 'Anyone who attempts a tensed practical judgement about what he has reason to do at a given time must accept (a) a tenseless practical judgement to the same effect about that time and (b) a belief about the relation between that time and the present which renders appropriate the particular tense employed.'[76]

It may seem that Nagel's argument is about consistency; that it claims that imprudent acts are inconsistent because they treat my current interests differently from my future interests. And it may also seem that we can negate this charge of inconsistency by insisting that there is a difference between my current self and my future selves; I am here *now* and they won't be here until the future. There is certainly some appeal to this thought, and the type of thinking it represents is familiar to us. We often weigh consequences for our future selves against immediate benefits and give more weight to the latter than the former, beyond that which could be justified solely by considerations of probability. Furthermore, it is far from clear that this type of thinking cannot be justified and regarded as rational. Thoughts such as, 'I'll regret this in the morning,' reflect our attitudes that, simplistically, pleasure now can be more important than pain tomorrow, and that, what is more, as we are the ones who will bear the consequences we have the authority to make the decision. To think like this is to collide directly with the second fundamental claim Nagel makes in his argument about prudence.

This claim is that we cannot escape the charge of inconsistency incurred by treating our future selves differently from our current selves without supposing that we are somehow separate from our future selves, and that to suppose this is to deny a

[75] *The Possibility of Altruism*, page 61.
[76] *The Possibility of Altruism*, page 68.

necessary aspect of our identities. This aspect is that we are 'temporally extended'[77] creatures: we do not just exist in time as all creatures do, but we are conscious of our existence in time. Nagel claims that, 'Those practical intuitions which acknowledge prudential reasons, and the motives connected with them, reflect an individual's conception of himself as a temporally persistent being: his ability to identify with past and future stages of himself and to regard them as forming a single life.'[78] For Nagel, then, if we claim that the interests of our future selves do not have the same influence on our reasons as the interests of our current selves then we deny our temporally extended natures, and, furthermore, deny the reality of our future selves. According to Nagel, this is a cost that we simply cannot accept: 'Any type of judgement which cannot be accommodated to that standpoint [of temporal neutrality] can be accepted only at the cost of dissociation from one's temporally extended self.'[79]

Of course, it is important to realise that, for Nagel's argument to be as free from dependence on subjective motivational states as he wants it to be, the cost of dissociation from one's temporally extended self cannot be motivational: it cannot simply be the discomfort we feel at the prospect. Rather, it must be a rational cost: temporal dissociation must be a violation of reason which is universally applicable to humans and sufficient to negate any apparent reasons which lead to it. The argument from the claims we have considered to such a violation of reason seems to proceed in six steps. First, reasons must be universal. Second, I seek a reason for action. Third, it is a given that my current interests are a source of reasons. Fourth, I am a temporally extended being. Fifth, my future interests are therefore as much a source of reasons as my current interests. Finally, whatever my reason for action is, it must take my future interests into consideration as well as my current interests. The fourth step is the crucial one as, according to Nagel, it represents a truth which we cannot ignore. We simply *are* temporally extended beings, and *are* aware of this, and to deny either is to lie about ourselves or to wilfully adopt a false belief. For Nagel, then, regardless of its motivational roots or consequences, reasoning which ignores our temporally extended nature or its implications is simply reasoning which has gone wrong. Consequently, Nagel's argument about prudential reasons is clearly an external reasons theory, as the reasons he identifies are dependent on something as apparently necessary as recognition of our temporally extended nature, and on nothing so contingent as an agent's subjective motivational states.

Nagel's argument for altruism has a similar structure to his argument for prudence. In this case, the relevantly similar entities which provide us with reasons are other human beings, the necessary aspect of our identity that we would be denying if we claimed that others were not relevantly similar entities is our perception of ourselves as one being among many others, and the cost of this denial is solipsism, or the denial of the reality of others. Just as with prudence, Nagel identifies two types of reasons, distinguishing those which are 'subjective' and contain a 'free agent variable' from those which are 'objective' and do not.[80] A free agent variable is a reference to an agent within a statement about reasons, where the value of this variable depends on who is making the

---

[77] *The Possibility of Altruism*, page 58.
[78] *The Possibility of Altruism*, page 58.
[79] *The Possibility of Altruism*, page 69.
[80] See *The Possibility of Altruism*, page 90.

statement. The most easily understandable examples of free agent variables are first-person pronouns. In the statement, '*I* have a reason to ϕ,' then the agent to whom *I* refers depends on who is making the statement: it is a free agent variable. This does not mean that objective reasons do not refer to agents at all. Rather, it means that when they do so they identify agents explicitly. In the statement, '*Agent A* has a reason to ϕ,' the reference to *agent A* is fixed. If the statement is true then it remains both true and concerned only with the reasons of agent A, whoever is making the statement. However, although objective reasons identify the agents to whom they apply, this does not mean that they can only ever refer to specific agents. Indeed, it is apparent that the sorts of reasons that Nagel is looking for are those which take the form, '*Everyone* has a reason to ϕ.' Again, if this statement is true it remains true and concerned with the reasons of everyone, regardless of who is making the statement. This means that, unsurprisingly, subjective and objective reasons bear the same relation to each other within Nagel's argument for altruism as tensed and tenseless reasons bear to each other within his argument for prudence. That is, Nagel is not denying that subjective reasons exist, or, indeed, that they are somehow inferior to objective reasons. Instead, he is arguing that in order to qualify as reasons, subjective reasons must effectively be localised versions of objective reasons.

It is this understanding of reasons as fundamentally objective, though permitting local, subjective manifestations, that allows Nagel to claim that, just as in his argument for prudence, the denial that others are relevantly similar entities to ourselves and are therefore sources of reasons has a cost, and the cost is solipsism. Without this understanding of reasons as necessarily objective, it would seem rather extravagant to claim that a denial or even just a lack of recognition of others as sources of reasons could be regarded as solipsism. After all, solipsism as we generally understand it seems to be a consciously adopted attitude to the world, rather than a mere implication of thoughtless actions. We can understand temporal dissociation as the result of a mere denial, but solipsism seems to requires us to consciously take at least one step further. However, if we understand reasons as necessarily objective, then by denying that other human beings are sources of reasons for us, we are also denying that they are sources of reasons at all. And by allowing that we are sources of reasons for ourselves, we are allowing that we are sources of objective reasons for everybody. Consequently, on Nagel's account, denying the reasons of others and acting for our own reasons is effectively to maintain that we are the only sources of reasons in the universe, and it is therefore legitimate to label us solipsists. While allowing that if we accept all of Nagel's claims this seems a justified conclusion to his argument, we should also note that the idea that solipsism is a result of denying the reasons of others is not as intuitively compelling as the equivalent argument that temporal dissociation is the result of denying prudential reasons; certainly it does not seem to offer the same intuitive threat to our identities. I believe that Nagel's argument for prudence provides a useful introduction to his argument for altruism not only because it is more straightforward, but also because it is a stronger argument, and sits more closely to our intuitions about the reasons that might be provided by our identities. However, we are less concerned with the success of Nagel's attempt to show the possibility and necessity of altruism specifically than with his success in showing the existence of external reasons generally, and his argument from identity to prudential reasons provides us with a real challenge, regardless of the problems inherent in using a similar argument to establish the existence of reasons for altruism.

In summary, then, for Nagel our identities as temporally extended inhabitants of a world which contains others like us have implications which provide us with reasons, the denial of which would violate our identities and therefore constitute errors in reasoning. There are aspects of this argument which are immediately unattractive: it lays the weight of the world on the agent's shoulders, and only makes incidental allowance for action on his or her own behalf. This burden can be alleviated by practical considerations, such as the claim that each agent is best placed to satisfy only a subset of the interests of the world, and that if every agent attempted to act on every reason created by every interest, such action would actually be counter-productive. However, the implication remains that all of us have reasons for action which go far beyond those which we acknowledge today, or even those which those of us living comfortable lives in wealthy industrialised nations already guiltily suspect that we have and are shirking. In later writing Nagel attempted to alleviate this unattractive aspect of his argument. In a note added to the beginning of *The Possibility of Altruism* he signals a shift in his views: 'This book defends the claim that only objective reasons are acceptable, and that subjective reasons are legitimate only if they can be derived from objective ones. I now think that the argument establishes a different conclusion: That there are objective reasons corresponding to all subjective ones . . . It remains possible that the original subjective reasons from which the others are generated retain some independent force and are not subsumed under them.'[81] Nagel's modified view is set out in his later book, *The View from Nowhere*.

In this book Nagel argues that although one of the distinctive things about human beings is that we can take up an objective standpoint, it is a standpoint that we have to take up: it is not the only or even the usual position from which we see the world. We also occupy a subjective standpoint, and some of our judgements, deliberation and reasons only make sense when we acknowledge the existence of this standpoint. The most obvious examples include those reasons and actions concerning people with which we have particular relationships, such as family and friends. As we saw, in *The Possibility of Altruism*, Nagel favoured the objective standpoint to the extent that he presented subjective reasons as mere local manifestations of objective reasons, and relegated the subjective standpoint to such a subservient position that he could make such claims as that, 'The only personal residue, therefore, which is not included in the system of impersonal beliefs to which I am committed by a personal judgement, is the basic personal premise itself, the premise which locates me in the world which has been impersonally described.'[82] In *The View from Nowhere* Nagel admits that he 'no longer thinks the argument works'[83] and this is borne out by his exploration of the two standpoints and his attempts to understand the tensions and relations between them. He also considers how the two standpoints can be reconciled, and what part attempts to reconcile them play in our lives. For example, in his discussion of freedom he considers the tension between our ability to regard the universe as a mechanism comprising an immutable series of causes and effects with no room for personal freedom, and our inescapable experience of just that freedom.[84]

However, allowing that the personal perspective plays a role in the explanation

---

[81] *The Possibility of Altruism*, page vi.
[82] *The Possibility of Altruism*, page 103.
[83] *The View from Nowhere*, page 159.
[84] See *The View from Nowhere*, section VII, 'Freedom'.

and justification of reasons does not make Nagel an internal reasons theorist: it just makes his external reasons theory more resilient. That he still believes that reasons exist independently of motivations is indicated when he says, 'The claim is that there are reasons for action, that we have to discover them instead of deriving them from our preexisting motivations – and that in this way we can acquire new motivations superior to the old.'[85] Indeed, although Nagel acknowledges the validity of the personal perspective in *The View from Nowhere*, his argument in this book can be regarded overall as strengthening his commitment to the preconception that reasons are universal. The influence exerted by the two perspectives could flow two ways: subjectivity could be taken to make objectivity into consensus rather than transcendence; or objectivity could be taken to provide a way of obligating agents who act on the whole for subjective reasons. It is evident that Nagel tends towards the latter, as indicated when he says, 'One translates one's own reasons into a form that can be accepted by people with different preferences, so that it can be used by anyone to account generally for his own reasons and those of others.'[86] So, although Nagel considers a broad range of topics in *The View from Nowhere*, the point that is of interest to us remains the same: that necessary aspects of human identity provide reasons independently of our subjective motivational states. In *The Possibility of Altruism* these aspects of identity are temporal extension and awareness of oneself as one among many similar others, while in *The View from Nowhere* they are our tendencies and capacities to adopt an objective standpoint.

2.1.2 Korsgaard's Argument from Identity

Before considering our response to the challenge posed by Nagel's arguments from identity to external reasons, we will consider another representative of this type of argument: that of Christine Korsgaard as presented in her book, *The Sources of Normativity*. Although Korsgaard's argument resembles Nagel's in some respects, in that it is an argument from identity, it differs because she approaches reasons from the opposite direction. As we have seen, Nagel takes as one of his starting points the principle that at least some reasons must be objective, and uses necessary aspects of human identity to determine what those reasons are. By contrast, Korsgaard uses conceptions of identity to argue that reasons are objective. We will not explore all of Korsgaard's argument here, as much of that argument is a distinctive interpretation of Kant's theories, and we will see much more of Kant in the next chapter when we discuss the derivation of external reasons from constraints imposed on principles of action. Rather, we will concentrate here on Korsgaard's arguments about identity, particularly those expressed in the third lecture in *The Sources of Normativity*, 'The authority of reflection.'

Korsgaard begins her argument with a claim which we can accept without difficulty: that human beings are self-conscious creatures. However, Korsgaard has a particular understanding of human self-consciousness as rational self-consciousness: that is, she takes our self-consciousness and capacity for reason to mean that we necessarily interpret the world and our place in it rationally. As long as we do not make too much of it, we could say that this argument presents us as standing one step away from the world,

[85] *The View from Nowhere*, page 139.
[86] *The View from Nowhere*, page 150.

with the distance between us and the world filled by reason. The importance of this rational remove for our discussion is that it means that for Korsgaard reasons do not only justify or explain action, but are required for rational creatures such as ourselves to be capable of action at all. So, Korsgaard claims that, 'If you had no normative conception of your identity, you could have no reasons for action, and because your consciousness is reflective, you could not then act at all,'[87] and that, 'The reflective mind cannot settle for perception and desire, not just as such. It needs a reason. Otherwise, at least as long as it reflects, it cannot commit itself or go forward.'[88]

Korsgaard's next step is to argue that operating as we do at this self-conscious, rational remove means that we must have conceptions of our identities. In other words, existence as a rational being means that I am not just somebody, but am a person with a conception of myself as a particular somebody. Initially, it seems that Korsgaard is concerned with more personal aspects of identity than the highly general aspects such as temporal extension and recognition of oneself as one among others which Nagel concentrated on. So, Korsgaard says that, 'The conception of one's identity in question here is not a theoretical one,' and that, 'It is better understood as a description under which you value yourself, a description under which you find your life to be worth living and your actions to be worth undertaking.'[89] She gives this conception the helpful label of *practical identity* and goes on to say that it, 'is a complex matter and for the average person there will be a jumble of such conceptions. You are a human being, a woman or a man, an adherent of a certain religion, a member of an ethnic group, a member of a certain profession, someone's lover or friend, and so on.'[90] Korsgaard claims that identities impose obligation and furthermore, that the primary source of obligation is the potential violation of identity: 'An obligation always takes the form of a reaction against the threat of a loss of identity.'[91] And just as we recognised the plausibility of the threats against identity introduced by Nagel, at least in the case of prudence, we can recognise the plausibility of the claim that identities impose obligations. The idea that there are things we just cannot do because of who we are is a familiar one; part of the reason that torturing someone into betraying friends and family is so terrible is that the forced betrayal is, among other things, a violation of identity. If the victim survives the torture he or she may never be quite the same person again, and not just because of the torture itself. The superficial plausibility of Korsgaard's account is also enhanced, especially from the standpoint of our account of internal reasons and of the individualistic intuition, by her apparent allowance that identity is personal and contingent, varying enormously between agents but nevertheless remaining of the utmost importance to those agents.

However, the very contingency of our various practical identities means that for Korsgaard they cannot be the ultimate sources of reasons. As she says, 'Because these conceptions are contingent, one or another of them may be shed. You may cease to think of yourself as a mother or a citizen or a Quaker, or, where the facts make that impossible, the conception may cease to have practical force: you may stop caring whether you live

[87] *The Sources of Normativity*, page 123.
[88] *The Sources of Normativity*, page 93.
[89] *The Sources of Normativity*, page 101.
[90] *The Sources of Normativity*, page 101.
[91] *The Sources of Normativity*, page 102.

up to the demands of a particular role.'[92] Just like Nagel, Korsgaard is looking for those aspects of identity which are general and inescapable; a practical identity which we must all necessarily possess. And, in true Kantian fashion, she argues in a reflexive move that the practical identity which is not contingent is our identity as beings who have a practical identity: 'What is not contingent is that you must be governed by *some* conception of your practical identity. For unless you are committed to some conception of your practical identity, you will lose your grip on yourself as having any reason to do one thing rather than another – and with it, your grip on yourself as having any reason to live and act at all.'[93] Korsgaard goes on to argue that this necessary commitment gives us a foundational practical identity as human beings, and furthermore, that it is this foundational practical identity which gives us the moral obligations that human beings are subject to: 'our identity as moral beings – as people who value themselves as human beings – stands behind our more particular practical identities. It is because we are human beings that we must act in the light of practical conceptions of our identity, and this means that their importance is partly derived from the importance of being human.'[94]

It is a long route from the necessary conception of practical identity and the foundational practical identity of human beings to individual, substantive reasons for action. But, if Korsgaard is correct in her argument, what she has done is show us how our identities as self-conscious, rational human beings place us on what we shall come to know as the Kantian conveyor belt. When we discuss Kant's theories in more detail in the next chapter, I will explain what I mean by the Kantian conveyor belt. Suffice it to say for the time being that it is the chain of reasoning that starts with the proposition that reasons require ultimate vindication and ends up with the Categorical Imperative. Korsgaard gets started on this chain of reasoning by, as we have seen, insisting that we must possess a fundamental practical identity and then asking what could possibly justify the nature of something so fundamental. This is enough to underpin external reasons and to present a challenge to our account of internal reasons. It is now time to attempt to meet that challenge.

2.1.3 Responding to the Challenge from Identity

The first thing that we must do when attempting to respond to the challenge presented by Nagel, Korsgaard and anyone adopting a similar line of argument is to avoid falling into a trap: we must resist the temptation to criticise their argument solely on the grounds of psychological plausibility. It is tempting to point out that no-one thinks in a way which explicitly articulates the considerations central to Nagel's and Korsgaard's theories. No-one stops to check whether the reality of other or future selves is being denied, or whether his or her foundational practical identity is being compromised before acting. However, although writers such as Nagel may make claims such as, 'I conceive ethics as a branch of psychology,'[95] it is evident that anyone adopting his argument or anything like it also conceives of psychology as subordinate to logical and metaphysical considerations. Neither Nagel's nor Korsgaard's theories require that the aspects of

---

[92] *The Sources of Normativity*, page 120.
[93] *The Sources of Normativity*, page 120-121.
[94] *The Sources of Normativity*, page 121.
[95] *The Possibility of Altruism*, page 3.

human identity they see as inescapable are acknowledged as such within the psychology of agents; they are only dependent on the existence and inescapability of these aspects, and their implications for the rationality of action. However, while this means that neither argument can be challenged on purely psychological grounds, it also means that they are limited in the appeal that they can make to psychological plausibility. While they do not depend on psychology, it would strengthen their case if they fitted our informal understanding and experience of human psychology. And there is an undoubted psychological appeal to the thought that assaults on identity provide us with reasons for action. However, if we examine it more closely we realise that much of this appeal has motivational roots on which writers such as Nagel and Korsgaard cannot rely for reasons. When we think of assaults on identity we think of those things which strike to our cores, such as being asked to betray loved ones, to deny fundamental beliefs or to compromise a trust. We imagine these assaults happening under extreme duress, such as torture, threats or the type of moral dilemma in which all courses of action seem as bad as one another. All of the passion and drama of such examples is of relevance only to arguments such as our account of internal reasons, which deal in individual psychology and motivations. None of it is available to those theories which deal only in the logical and metaphysical implications of identity. Of course, neither Nagel nor Korsgaard attempt to gain any spurious credibility by appealing to such psychological considerations, but we must be aware that simply citing the prospect of assaults on identity raises such considerations.

Once we set questions of psychology aside, we can see that what is essential to arguments from identity to external reasons is the relationship between the implications of the supposedly inescapable aspects of human identity and the implications of actions. Nagel, Korsgaard and anyone adopting a similar line of argument effectively claim that anyone who believes that he or she has a reason to perform an action whose implications contradict the implications of identity has simply made an error of reasoning. For Nagel, it is a fact that we have identities as temporally extended beings, and any action which treats our future selves as if they are not sources of reasons equivalent to our current selves is based on a factual error, just as any action would be if based on some other fundamentally erroneous assumption, such as that time sometimes runs backwards, or that the world came to an end yesterday. This pared down understanding of the argument provides us with our first response to this type of argument: in order to determine practical reasons, human identity would have to have a limited, determinate set of implications, and our experience is that this is not the case.

If we allow that human identity has implications pertinent to action, we must also allow that it has rather more than those identified by Korsgaard and Nagel. We cannot exhaust all of the implications of identity here, but we can note that there are implications which contradict those identified by Nagel and Korsgaard and which, more generally, contradict any attempts to derive universal reasons. The strongest of these is that each of us has an identity as an individual with our own interests, and that the influence of these interests on us is not just because we are the people who happen to be best placed to satisfy them, but because they are *ours*. Giving up this relationship to our interests and regarding ourselves as agents acting on behalf of all of the interests in this world is at least as much an assault on a fundamental and inescapable aspect of our identities as denying that we are temporally extended. It is rare for anybody to genuinely attempt to adopt the viewpoint in which their interests count no more than anybody else's, and even

rarer for anyone to achieve it.[96] We would most naturally describe such a person as self-denying, not just in the sense that that person is denying his or her interests, but also in the sense that that person is denying his or her identity. This recognition that our identities necessarily contain acknowledgement of our own existence and a unique relationship with our own interests may carry connotations of selfishness or egoism. However, I hope that we have seen enough in earlier discussions of motivations to realise that such thoughts do not have to lead in this direction: the interests and motivations of individuals can lead to behaviour which is self-sacrificing as well as self-serving.

The potentially contradictory implications of identity are even more apparent when we consider that the identities of individual agents do not simply comprise the common aspects we have considered, such as the recognition of ourselves as entities with our own interests, but also include elements peculiar to the individual. As we have been concentrating on the common aspects of human identity we have not considered these elements in any detail, but once we do consider them we realise that they are a fundamental part of our understanding of identity. When we think of an individual we think of his or her identity as that individual first, rather than his or her identity as a member of the human race. These individual aspects of identity vary enormously between individuals and naturally carry their own implications, at least some of which are likely to contradict the implications of those common aspects of identity we have considered so far. As we have seen, Korsgaard claims that as these individual elements of identity are contingent and could be theoretically discarded, what really matters are those remaining elements which could not be discarded: our identities as agents who need to operate under some form of practical identity. We can respond to the claim by pointing out that even while these individual elements could theoretically be discarded, in the sense that we can imagine an agent who no longer possessed them but still possessed an identity as a rational, human agent, to discard many of these elements would in itself constitute a devastating assault on the agent's identity. Indeed, an assault which left nothing but the common aspects of rational, human identity would be so devastating that it would not make sense to say that individual elements of identity have been discarded: we would instead say that the agent had lost his or her identity. Although they are contingent and vary between individuals, agents possess elements of identity which are as fundamental to them and whose implications are as inescapable as those elements of identity which are common to all humans. If we acknowledge the presence and influence of these additional, contingent aspects of identity, it no longer seems likely that any common human elements of identity and their implications will produce substantive, determinate conclusions about action which are not contradicted by other aspects of our identities. What seems more appropriate, given the more complex picture of identity which appears

[96] Of course, there are some religions in which the loss of personal interests is a definite goal, most notably some forms of Buddhism. However, it is worth noting that, at least in the case of Buddhism, the goal is not just to lose awareness of one's own interests, but of *all* interests: the person aspiring to enlightenment does not seek to fully recognise the reality of others, but to deny the reality of everybody, including him or herself, thereby extinguishing his or her existence. It is also worth noting that the popular forms of such religion contain all the usual manifestations of followers seeking supernatural assistance in achieving material goals, and that spiritual tourists from other cultures are often pursuing distinctly personal aims such as self-awareness and self-actualisation. For a brief discussion of the development of Buddhism from a quest for self-extinction to a more popularly palatable vehicle for salvation see Edward Conze, *A Short History of Buddhism.*

to be natural, is a model such as our account of internal reasons, in which practical conclusions are drawn out by deliberation from a mass of potentially contradictory considerations, and where essential aspects of human identity constitute, at most, some of these considerations.

We may wonder why, if this picture of human identity as complex and indeterminate seems so natural to us, Nagel and Korsgaard have ended with competing pictures which diverge from it so strongly, and an understanding of reasons which is apparently derived from these pictures. I believe that the answer is that their positions have actually developed in the opposite direction: they have started with the assumption that reasons must be universal and have developed their conceptions of identity from that assumption. This is most apparent in Nagel. As we have seen, he derives implications of identity from our temporal extension and status as individuals among similar others. And we certainly cannot deny that we are temporally extended or that we are individuals among other individuals: to do so would be to go wrong in reasoning. However, it is not the same thing to recognise the existence of these other entities as it is to recognise them as sources of reasons, let alone to recognise them as sources of reasons for an agent which are equivalent to that agent's own interests. The only basis which Nagel has for supposing that it is the same thing is the conception of reasons he starts out with; that reasons are necessarily objective, general and universal, and that the only relevance of individual, local circumstances is that they determine how these general reasons must be applied.[97] But we have been given no justification, other than Nagel's assertion, for understanding reasons on this basis. Our own exploration of reasons has given us a plausible model of reasons and actions, in which reasons are rather less universal and more indeterminate and contingent than would be allowed by Nagel. Even if this model was shown to be wrong, its plausibility at least means that alternative models such as Nagel's require justification. So, it appears that Nagel has not discovered aspects of identity whose implications are that there are necessarily reasons independent of motivations, but that he has rather discovered aspects of identity which tell us what those reasons might be if his conception of reasons was right. We can make this into a more general point by observing that this flaw in Nagel's argument indicates that the implications of identity are limited to just that: what identity can imply. Taken alone, these implications cannot give us independent general criteria for what counts as a reason or as sound deliberation; we must discover that through the consideration of reasons rather than identity.

Korsgaard might seem to be in a better position because she supposedly argues for the universal nature of reasons rather than assuming it at the outset. However, Korsgaard's conception of identity still seems alien to us, and this is because it is inexorably foundational; it depends on the assertion that much of what informs our practical lives, and what we would normally regard as part of our identities, can be discarded or disregarded as we seek the ultimate practical identity which is the source of reasons. The provisional identities which Korsgaard claims that we can discard seem much closer to our understanding of identity than the abstract conception of ourselves as beings who require practical identities. The description of an agent as a lawyer or a doctor or a parent seems much more plausible and accurate than the description of an agent as a being-who-must-possess-a-practical-identity who is acting as a lawyer, as a

[97] For example, see references 79 and 80 above.

doctor and so on. This becomes even more apparent when we realise that what we consider to be our most fundamental identities are even deeper and more personal than these generic roles. What lies beneath the role of doctor or lawyer is the identity of the agent as him or herself, even though this identity remains contingent on the agent's individual make-up and circumstances. Indeed, the conception of the underlying fundamental practical identity reveals that Korsgaard's basic assumption about the nature of identity depends on a preconception about the nature of reasons just as strong as Nagel's. She is supposing that our identities must be underpinned in some fashion; that they must have some sort of general conceptual backing. But our understanding of identity does not require this backing unless we are starting from the assumption that we are seeking the basis of universal reasons. If we do not start from this assumption then, while we may still talk of our fundamental identities we will mean those things we could not abandon without losing ourselves, and will suppose that these are peculiar to individuals rather than anything so abstract as the identity of a being who must adopt a practical identity.

So, while we can acknowledge that our identities are powerful influences on our reasons and our actions, neither considerations of necessary, common aspects of human identity, nor the implications of action seem enough to provide us with substantive, determinate reasons which apply to agents regardless of their motivations, unless we bring with us question-begging conceptions of the nature of reasons. The argument from identity is not a challenge to our account of internal reasons as long as it depends on the logical and metaphysical implications of identity. These implications are simply further facts about us, some of which will figure in deliberation and some of which won't. These facts are not of the sort which can be contradicted by action or deliberation which does not directly imply such contradiction, or pass very close to it. In the absence of motivations these further facts will come up against the same objection that they will always face from Humean theories: we fail to see how they can explain action on their own. Furthermore, we have no reason to believe that the implications of common aspects of identity should necessarily be more influential than the implications of the aspects of identity which vary from individual to individual but which are nevertheless fundamental to those individuals. The real plausibility of the claim that aspects of identity produce reasons comes when we allow motivations to play their part. We can then adopt an understanding of identity in which a starting set of motivations, along with those which are settled by experience and deliberation, give us a powerful explanation of why some reasons have the hold over us that they do. And this is, of course, a conception of identity which is entirely compatible with the internal reasons account.

2.2 Principles

The best representative of the argument that universal reasons can be discerned through the principles on which we act can be found in the work of Immanuel Kant, particularly in the *Groundwork of the Metaphysics of Morals* and the *Critique of Practical Reason.* In considering Kant's arguments we must be aware of the relationship they have with earlier empiricist arguments to which they provide a response and a challenge. Although Kant arrives at conclusions which are far removed from empiricism, one of his starting points is that the empiricists who came before him, and particularly Hume, have set him a problem: they have identified the ravenous demands that rational enquiry makes on beliefs and evidence, and have argued that these demands can never be met. In the *Critique of Pure Reason* Kant lamented that John Locke, 'opened a wide door to extravagance (for if reason is once allowed right on its side, it will not allow itself to be confined, by vague recommendations of moderation)' while David Hume, 'gave himself up entirely to scepticism having, as he believed, discovered that what passes for reason is nothing but a pervasive illusion of our knowledge faculty,' and proposed to, 'make a trial whether it be not possible safely to conduct reason between these two rocks, to assign her determinate limits, and yet leave open for her the entire sphere of legitimate activity.'[98] Furthermore, Kant was not amenable to what Hume saw as the only sensible reaction to his apparently sceptical conclusions: to accept that despite a lack of foundations which can withstand the insatiable demands of rational justification, we are nevertheless capable of going on living, knowing, judging and acting. For Kant, those unassailable foundations are not only desirable but necessary.

Kant's attempt to find vindication in such foundations results in the construction of what I have flippantly referred to as the Kantian conveyor belt; that is, a line of reasoning which, if we can find a point of entry to it, leads us inexorably to the conclusion that we are governed by Kant's famous Categorical Imperative. The Categorical Imperative appears in Kant's work in several forms, but the one which is best known and which provides us with the clearest basis for discussion is that known as the Formula of Universal Law: 'act only in accordance with that maxim which you can at the same time will that it becomes a universal law.'[99] The conveyor belt proceeds in four stages. The first stage is to suppose that actions are expressions of principles. At this stage Kant does not need us to suppose that these are general principles, or even that they can be justified, but he does require us to acknowledge that, because we are rational beings, when we act there is something in addition to the act itself; there is the principle upon which we act. So, on this account, any act committed by a rational agent, no matter how thoughtless or casual, is an act on a principle, even if that principle is selfish or venal. The second stage of the Kantian conveyor belt is to suppose that, even though many of the principles on which we act could not be justified, as principles they stand in need of rational justification. The third stage is to recognise that, as discussed by the empiricists, we cannot appeal to anything in the world that can provide this justification; there is nothing that reason will accept as a foundation. Or almost nothing. In the fourth stage of the Kantian conveyor belt we make what Kant refers to as the 'Copernican

[98] *Critique of Pure Reason*, page 97.
[99] *Groundwork*, 4:421.

shift'[100] and radically change our perspective. If there is nothing external which can be appealed to to provide rational justification, then we must look for it within the requirement for justification itself. In other words, the way in which we find principles which stand up to the insatiable demands of reason is not by continuously looking for foundations which will inevitably be undermined, but by looking directly for those principles which meet the demands of reason. In Kant's scheme, such principles have the status of universal law because they can always be justified for all agents, and they acquire this status just because they can be conceived as universal law. We have reached Kant's famous Categorical Imperative in the form of the Formula of Universal Law: 'act only in accordance with that maxim which you can at the same time will that it becomes a universal law.'[101] This argument can be a little dizzying: Kant is accepting the Humean conclusion that reason sweeps away the ground under our feet by denying the solidity of any empirical foundations, but stops us from falling by finding justification within reason itself. However, although this argument can be bewildering and difficult to grasp, once it is grasped it is curiously compelling. I have adopted the metaphor of the Kantian conveyor belt because once we enter Kant's chain of reasoning it propels us inexorably towards the Categorical Imperative.

And, of course, if we allow ourselves to be conveyed to this conclusion regarding reasons for action, then we must abandon our account of internal reasons. Kant's argument implies not only that reasons exist independently of motivations, but that all reasons *must* be independent of motivations. Motivations are precisely the form of empirical consideration that Kant rules out as capable of providing justification. So, if we wish to retain our attractive, plausible position, we must respond to the challenge set by Kant. There are many possible ways to make such a response, some of which are purely based on the superficial unattractiveness of Kant's austere and demanding account. However, such responses do not engage with the heart of Kant's theory, and to rely on them would be to fall into a similar trap as those external reasons theorists who attack positions resembling the sub-Humean model, without acknowledging the breadth and complexity of the actual Humean position and its derivatives. I believe that to respond adequately to Kant we must not just show that the Categorical Imperative produces unattractive or counter-intuitive conclusions, but rather that we do not need to get on the Kantian conveyor belt in the first place. We will be helped in this attempt by a modern interpreter of Kant: Onora O'Neill, particularly her collection of essays *Constructions of Reason.*

O'Neill is not only a well-known and lucid interpreter of Kant's sometimes impenetrable work, but is also a perceptive critic of those supposed Kantians who discard, contradict or ignore the less palatable aspects of his theories, without which his arguments lose much of their elegant and compelling nature.[102] By modifying his theories, these supposed Kantians make him less interesting to us; by accepting something like the Humean theory of motivation they make him into an internal reasons theorist. Within O'Neill's interpretation we get a comprehensible version of Kant's unyielding focus on principles and reason; we get the external reasons theorist who

---

[100] See *Critique of Pure Reason*, page 17.
[101] *Groundwork*, 4:421.
[102] For example, see 'Autonomy: The Emperor's New Clothes,' the inaugural address to the joint session of the Aristotelian Society and the Mind Association, 2003.

presents a real and serious challenge to our account. O'Neill also extensively discusses a concept which she takes to be central to Kant's practical philosophy, and which I believe represents both the entry point to the Kantian conveyor belt and the most promising point for us to challenge: the concept of the underlying principle of action, or the *maxim*.

It is easy to overlook that the Categorical Imperative, when expressed as the Formula of Universal Law, refers not only to action, will and law, but that it uses the concept of the maxim to link them together. This concept may seem rather mundane and workmanlike, particularly in the company of lofty notions found within Kant's theory such as the supreme principle of practical reason, rational autonomy and universal law, but, as O'Neill says, 'contrary to appearances, this is not a trivial part of his criterion of morally acceptable action.'[103] I also believe that it is the point at which, at least in the context of our discussion, Kant's theory should be challenged. Maxims stand as intermediaries between our individual, specific actions, and the obligations Kant supposes to arise from our rational natures. They serve three purposes within his scheme. The first of these seems relatively innocuous. As we have seen, Kant supposes that all action by rational beings is action on some principle, and maxims capture the underlying principle of actions. As we have mentioned, at this stage we need not suppose that these principles could qualify as universal law, or even that they are general; they could be entirely subjective. However, as we have also seen, acknowledging that all action is action on a principle has significant implications; that principle, and by extension that action, becomes subject to the demand for justification. So, the second role of maxims is as the vehicle between individual actions and the Categorical Imperative; capturing the principle of action is necessary to submit that action to the test of the Categorical Imperative; and it is also what makes these actions subject to this test in the first place. However, we do not have to wait for the occasion of a particular action to test the maxim of that action. The third role of maxims in Kant's scheme is that they provide a means of testing whether certain general types of action can be justified, and thereby a means of deriving general rules about whether those types of action are permissible, required or forbidden. Thinking of such maxims outside the context of particular actions is not something that Kant's scheme demands or expects that we do; but it is an activity for moral philosophers, and is the source of Kant's more hotly contested examples of what is and is not morally permissible,[104] as well as the perception of the Kantian scheme as tightly bound by rules. So, given these different roles which maxims play in Kant's scheme, when discussing maxims we must always remember that they fall into three nested categories which correspond to these roles: the category of untested maxims, which will have varying levels of subjectivity and objectivity, and which may or may not be capable of passing the test of the Categorical Imperative; the smaller category of those maxims which pass that test, and which therefore constitute moral obligations; and the even smaller category of maxims which have been tested by theorists such as Kant outside the context of individual actions, and which have been discovered to constitute general moral rules.

The different levels at which maxims operate and the purposes they serve in Kant's scheme become clearer when illustrated with an example. Imagine that a couple I

---

[103] 'Consistency in Action' in *Constructions of Reason*, page 83.

[104] Perhaps the most notorious example is Kant's insistence that the truth should be told even if a lie could save someone from being murdered. See *On a Supposed Right to Lie from Philanthropy*, 8:247.

meet on holiday ask me to dinner once we have returned, but I realise that I don't really want to keep up the relationship. At the same time I don't want to hurt their feelings, so I make various excuses about my busy schedule and prior commitments in the hope that they will gradually lose interest without feeling slighted. These excuses are not based in truth, so I am lying. There are many possible maxims which could be used to capture the principle of my action (although we will also ask later whether it is possible that I am not acting on a principle at all) but let us suppose that my actual maxim is, 'I shall lie whenever I need to get out of a socially awkward situation.' This maxim cannot but start by belonging to the category of untested maxims. However, it is rather less certain that it belongs to the category of maxims which pass the test of the Categorical Imperative. It can be difficult to determine whether a maxim passes this test, as it is not always immediately obvious which maxims could be willed as universal law. There are two standard ways in which the test can be failed: because willing the maxim as universal law involves a conceptual contradiction, as would be the case if we willed the maxim that we could break promises whenever convenient, as such a maxim would contradict the meaning of promises; or because willing the maxim as universal law involves a practical contradiction, as would be the case if, through selfishness, we willed maxims of neglect or inaction, which would serve to frustrate the fulfilment of our own selfish ends if established as universal law. Kant claims that any maxim involving lying fails the test in the first way; it involves a conceptual contradiction, because the practice of truth-telling, on which lying depends, would be rendered meaningless by the universal permissibility of lying.[105] This means that, while the proposed maxim may be the genuine principle of my action, my action would not be permissible under Kant's scheme, as it fails the test of the Categorical Imperative. This is all that is needed within Kant's scheme to deliver substantive moral conclusions regarding particular instances of action. However, the scheme does not need to stop there; hence the third category of maxims. It is possible to generalise the maxim further, and by considering whether this maxim is capable of passing the test of the Categorical Imperative, to produce general moral rules governing a broad range of situations and maxims. The more general maxim which can be derived from the specific maxim we have been considering is, 'I shall lie when I can gain some advantage by doing so,' and is a maxim to which Kant considered directly and found to be impermissible.[106]

Outlining these categories of maxims, from those which capture the subjective principles of action to those which constitute general moral guidelines, helps us to remember which sort of maxim we are dealing with at each part of our discussion. According to Kant's scheme, we always act on maxims, we should always act on maxims which have been subjected to the test of the Categorical Imperative and which have passed that test, and we have a better chance of acting correctly if we follow those general rules which can be derived by submitting highly general principles to the test. However, we can challenge the progression through these categories of maxims, and up

---

[105] The conclusion may seem counter-intuitive, as it superficially seems neither harmful nor logically incoherent to allow that people may tell white lies from time to time, especially to protect the feelings of others. However, because we are forced under Kant's scheme to consider maxims as possible universal laws, we put in the position where the most innocuous and well-meaning lie is an assault on our commitment to truth and the practice of speaking truthfully. For a useful discussion of this point see Roger J. Sullivan's *Immanuel Kant's Moral Theory,* pages 170-173.

[106] See *The Metaphysics of Morals*, 6:481.

the Kantian conveyor belt, by questioning whether the assumption at its very beginning is true; by asking whether we always act on maxims.

2.2.1 Challenging Maxims

We will challenge the concept of maxims by attempting to show that there are at least some rational actions which cannot be made sense of as the embodiment of a principle, and that therefore the criteria of justification applied to principles do not necessarily have to be applied to all rational actions. In accordance with our discussion of the different categories of maxims, we should note that what we are challenging here is the claim that we always act on maxims, whether those maxims are capable of passing the test of the Categorical Imperative or not. In arguing for this case we will make life easy on ourselves by considering a category of action which is least amenable to interpretation as acting on a principle: action which I shall label 'trivial'. This label is only a tool, and I do not wish to make too much of it, as I hope that it is obvious what it means: it refers to those actions which could as readily be performed as not performed, without making much difference to the agent. A paradigm example of trivial action would be that of taking a break from writing this passage to go and get a cup of coffee, when I am not particularly hungry, thirsty or in need of stimulation. The thought simply occurs to me that a cup of coffee might be nice about now, and if I have one it is quite nice, but if I don't then I carry on working without it. Insipid adjectives such as 'nice' and the thought that my action might be habit as much as desire are indicators of trivial action.

We can begin our challenge by making the informal observation that it simply seems absurd to say that trivial action is action on a principle; our ordinary intuitive understanding of our behaviour is that we sometimes act with little or no thought about what we are doing, and still less about any possible principle on which we could be acting. To insist that such action is action on a maxim seems to attempt to force part of our common experience of action into a preconceived theoretical understanding of what action should be. Yet just such insistence is made by Kant and his followers; Onora O'Neill claims that, 'Even routine or thoughtless action is action on some maxim.' [107] To deal with such insistence we must attempt to articulate our intuition about the absurdity of trivial action being action on a maxim by considering what such a maxim would have to be like to satisfy both Kant's scheme and the nature of the action itself.

The requirement imposed by Kant's scheme is that maxims must be of a form which is capable of being submitted to the test of the Categorical Imperative. Once again, this does not mean that they must be capable of passing this test, even though Kant maintains that these are the only maxims we should act on; all that is required at this stage is that the test can be attempted. This means that the maxim of any action must have at least two characteristics. Firstly, it must be capable of being formulated as a principle. This means that we cannot simply determine that an action appears to express some sort of vague, hazy attitude or belief; it must be conceivable that some fully articulated principle could be wrung from the agent's action, even if this is very hard in practice. The second characteristic derived from the Categorical Imperative is that whatever principle is finally extracted from the action, that principle must be capable of being *willed* by the

[107] 'Consistency in Action', page 84.

agent; that is, it must be possible for the agent to recognise the maxim as correctly describing his or her action, to rationally endorse it, and to adopt it as the principle of his or her action. The test of the Categorical Imperative requires, of course, that it must be capable of being willed as universal law but, yet again, we are concerned for the time being with what is necessary to submit the maxim to this test, not what is needed to pass it. The important point here is that, to qualify as a maxim, the principle underlying the agent's action cannot stand apart from the agent's will; to be the principle of the agent's action it must be willed, or at least willable. It is the willing of maxims, as well as their status as principles, which makes them subject to the demand for rational justification; on Kant's understanding, by willing maxims, we assert at least, 'This action is allowed,' and such assertions demand justification. To act on a maxim cannot be merely to behave consistently with the description of one's action under a principle; it must be to will that principle.

It is slightly more difficult to identify the characteristics which the maxim of a trivial action such as going for a coffee break must have from the nature of that action, because the boundaries and definition of the action itself are hazily defined, as they often are for trivial actions. However, I think that we can discern two such characteristics. Firstly, my maxim is not psychologically articulated. It would be surprising if, in taking a coffee break, I consciously formed a principle concerning the taking of coffee breaks. It is most likely that I will barely deliberate at all before taking such action; indeed, if I am particularly distracted, or particularly deep in thought, I may glance at my empty mug and find myself in the kitchen with the kettle on before I properly realise what I am doing. So, for such trivial actions, the maxim of my action must be such that it does not need to consciously register in my mind at all. The second characteristic of the maxim of trivial action which we can derive from the nature and circumstances of that action is that the adoption of such maxims must be amenable to a cheerful inconsistency. Actions which are trivial do not matter to me very much, and acting in the same way in comparable circumstances matters barely at all. So if I take a break to get a cup of coffee now, it does not imply that I will do the same thing tomorrow when I am in relevantly similar circumstances. Such inconsistency does not bother me at all, and we would consider someone very strange who challenged me to explain myself with respect to it.

So, now we must ask whether these sets of characteristics are compatible with each other. That is, is it possible to construct a maxim which is compatible with the Kant's theory and which is compatible with the nature and circumstances of trivial action? A case could be made for saying that the two sets of characteristics are compatible, at least superficially. Kant's theory does not require that maxims are consciously articulated if they are to figure in the agent's psychology. One of Kant's gloomier conclusions about human behaviour is that we can never be certain which maxims we are acting on, even if we believe ourselves to have consciously chosen our maxims.[108] The person who is apparently acting on maxims of charity may actually be acting on maxims of pure self-interest, and may have deceived him or herself as well as any observer. Similarly, actions which are characterised by cheerful inconsistency could be seen as compatible with Kant's theory due to the agent's inconsistent adoption of maxims. Kant need not suppose that agents always act on the same maxims, even when they are performing actions which are outwardly similar: all that is claimed is that agents

[108] *Groundwork*, 4:407-8.

always act on some maxim. The interpretation even allows that agents could act on different maxims in equivalent circumstances yet still be judged rational and moral as long as those different maxims pass the test of the Categorical Imperative.

However, despite this superficial compatibility, problems arise as soon as we look deeper, particularly when we consider how Kant's theory must deal with the thought that maxims must be *willable*; as mentioned earlier, the agent must be able to recognise, endorse and adopt the maxim as a principle of action. The requirement that a maxim is willable does not just mean that maxims must be of a form which could be willed, but that they can actually be willed by the agent who possesses them. In other words, for a maxim to qualify as an agent's maxim of action, the agent must have a psychological commitment to that maxim at some level. It cannot simply be a principle which gives a convincing or plausible description of the action; it must be the maxim of that agent and of that action. Furthermore, this commitment cannot be to the execution of a particular action, or even to the ends that action will produce, but to the *principle* of that action. And, because maxims are ubiquitous, the principles to which agents have some form of commitment must include the principles concerning trivial action. We could come up with many principles which were superficially compatible with a specific instance of trivial action, but it is extremely difficult to imagine psychologically plausible principles which agents would recognise, endorse and adopt as the actual basis for those actions. Remember that we are not talking about commitment to an action, or even the ends of that action, but to the principle under which the action is conducted. It is not impossible that when I go and get a cup of coffee my mind contains a commitment to a principle of action governing going to get cups of coffee, but it is implausible; for such trivial, habitual actions we are stretching our terms if we say that we are even committed to the action, let alone the principle of the action.

So, if we want to apply the concept of maxims to cases of trivial action, it appears that we not only have to accept the plausible claim that principles of action are not always fully articulated, but that we also have to accept the rather less plausible claim that principles which we cannot imagine as commitments nevertheless exist as commitments within our psychologies in a form which we cannot discern. We can neither afford nor need to be so generous, and can conclude that, on the basis of our discussions so far, trivial action is not action on a maxim. There are, of course, ways in which Kantians could defend the concept of maxims from this challenge. Three obvious defences present themselves. Firstly, it could be claimed that we have made life rather too easy for ourselves by dealing only with trivial actions, and that actions which are more significant fit the concept of maxims rather better. So, if I am considering jumping off a bridge, then the principle of my action may be rather more clear and rather more prominent in my psychology than if I am considering getting a cup of coffee. Although this objection does force us to acknowledge that there are situations in which we do seem to be deliberating and acting in accordance with maxims, it has problems of its own. To start with, it is not enough to show that *some* actions are conducted on the basis of articulated principles. The point of the ubiquity of maxims is that they apply to all of our actions, meaning that we can never avoid getting on the Kantian conveyor belt. As we have seen, O'Neill insists that, 'Even routine or thoughtless or indecisive action is action on some maxim.'[109] So, we may have made life easy for ourselves by picking the category of

[109] 'Consistency in Action', page 84.

action least suitable to maxims, but we can do this because the theory requires that maxims are ubiquitous.

Furthermore, part of the reason that we chose to talk about trivial actions is that, as well as testing the concept of maxims, we can discuss them without introducing additional considerations, such as those concerning morality and relationships, which complicate the argument. However, when we do consider more significant actions, we find that they are not so different from trivial actions. Consider an example used by Bernard Williams in his paper 'Persons, Character and Morality.'[110] In this paper, Williams imagines a situation in which a man has the opportunity to save several people from drowning, one of whom is his wife and the rest of whom are strangers. Unsurprisingly, he saves his wife; and there are few of us who would challenge him for doing so. The point of Williams' example is that it seems odd to say that the man ought to deliberate to the conclusion that he is justified in saving his wife before actually saving her; and just as odd to say that if he does not do so his action is somehow less than rational.[111] We can use this example in our current argument by considering whether we should understand the man's action as being governed by a maxim which figures somewhere in his psychology. If so, the maxim is certainly not consciously articulated; the deliberation that Williams plausibly claims is unnecessary would be needed to identify and articulate that maxim. However, as well as being unarticulated we can also see a maxim in this situation as unnecessary and inappropriate: we can better understand the agent as committed to the ends of his action, or simply as committed to his wife than as committed to a principle. In at least this case of significant action it seems a poor fit to attempt to describe this action in terms of maxims at all.

The second defence of the claim that even trivial action is action on a maxim is that we have made a mistake which O'Neill warns against: we have conceived maxims as precise algorithms governing a specific action rather than as broader principles governing a range of action. So, on this defence, the details of trivial action may not be specified in a particular, detailed maxim but rather covered in the broader terms of a more general maxim. I think that we have avoided this mistake as we have not attempted to formulate maxims of specific trivial actions but have rather considered what characteristics such maxims must have. However, it may be that maxims are so general that the characteristics we derived from the nature of such trivial actions, particularly that they are compatible with cheerful inconsistency, are not actually necessary characteristics at all: trivial actions are so far beneath their attention that they are not influenced by them. However, conceiving of maxims in such a way would leave us with a problem. If the intention is to avoid the conclusion that some actions should not be regarded as actions on maxims then the maxim would have to be sufficiently general to avoid acquiring characteristics determined by the nature of the actions with which it is concerned, yet sufficiently related to such action to be considered the maxim of the action.

Such maxims are not inconceivable. For example, the action of going to get a cup of coffee while writing a thesis may be action on the maxim, 'Keep yourself refreshed when engaged in intellectual activity,' or, 'Don't work so hard that you forget your bodily needs,' or, 'Give yourself time to reflect on your arguments,' or even, 'Make sure

---

[110] See 'Persons, Character and Morality' in *Moral Luck*..

[111] We should also note that the question here is not whether he has time in this crisis situation to conduct such deliberation, but whether it is necessary or appropriate for him to conduct such deliberation at all.

that you use up all the groceries you have bought before their expiry dates.' Indeed, the problem here is not that we cannot conceive of one, two or a multitude of maxims under which the action could be adequately described, but that we have difficulty in supposing that any or all of these should be understood as expressing a commitment of the agent to a principle. For, even though maxims are supposed to be sufficiently general that they do not specify detailed actions, they are also supposed to be maxims *of* those actions. And, even if we are able to find some agents some of whose trivial actions do express allegiance to general principles of the type we have just considered, it does not seem plausible to suppose that all such actions express a similar allegiance. This becomes apparent when we realise that, unfettered by considerations such as the actual beliefs and attitudes of the agent we could come up with a potentially infinite range of maxims which are apparently compatible with trivial action, especially when we consider the negative implications of the agent's action. So, we could imagine maxims such as, 'When thirsty, drink in preference to working,' or, 'When thirsty, drink in preference to phoning your mother,' or, 'When thirsty, drink in preference to doing charity work.' Yet we have no grounds for supposing that such principles which the action conceivably *could* express are the principles which the action *does* express, or, as we are arguing, that the action expresses any principles at all. Ensuring that we conceive of maxims as broad principles rather than as detailed algorithms does not seem to make the fit between the concept of maxims and our experience of trivial actions any better.

The third defence of the concept of maxims is to argue that we have used an understanding of maxims which places too great an emphasis on their presence in the agent's psychology. The idea behind this defence is most clearly expressed by O'Neill in the postscript to her paper 'Kant After Virtue.'[112] In this paper O'Neill originally offered the corrective to criticisms of Kant which we have already seen: she responded to Alasdair MacIntyre's claims in *After Virtue* that Kant presents an unrecognisable and unliveable picture of moral life by treating morality as the prescription of narrow rules[113] by arguing that this understanding of Kant is mistaken because maxims are general principles of action rather than detailed algorithms. She even goes so far as to suggest that a sufficiently general understanding of maxims allows us to understand them as something akin to the virtues; an understanding that obviously has implications for the status of maxims as psychological contents of the agent. In the postscript, however, added when the paper was published in the collection *Constructions of Reason*, O'Neill disowns this conclusion and its implications. Here she says, 'The point about moral principles on Kant's account is not that they are psychologically inward – to think of it in that way, even if one allows for the opacity of maxims, is to adopt a "Cartesian" view of maxims as agents' mental states. Maxims may be inferences from action; they may be imputed to practices or to institutions rather than to individuals . . . The sense in which maxims are inward is only that they are not outward – they are not inscribed on the surface of action. It would be better to describe them as *underlying* rather than as psychologically inward principles.'[114]

This argument may offer a defence against our challenge to maxims because that challenge relied in part on the implausibility of any agent holding a commitment to the

---

[112] See 'Kant After Virtue' in *Constructions of Reason*.
[113] *After Virtue*, pages 43-47.
[114] 'Kant After Virtue', pages 161-162.

maxim of a trivial action, assuming that such a maxim could be formulated in the first place. On O'Neill's interpretation it seems that we need not suppose that the agent holds this commitment, making the existence of maxims of trivial action rather more plausible. However, this defence has problems of its own. Firstly, even though it may remove the need for agents to commit psychologically to the maxim of a trivial action, it still requires that the maxims of such actions can be formulated, and this seems as difficult as ever: even if we can come up with countless maxims which could fit the agent's action, it is rather harder to find the actual maxim on which the agent did act, if he or she acted on a maxim at all. Underlying intentions may be obscure to agents themselves, let alone others, and circumstances may prevent us from ever knowing what these intentions were. Psychotherapy notoriously takes years to uncover the roots of behaviour, and even then it may be uncertain that this is really what has been done. In a starker and regrettably topical example, we are unlikely to ever know the underlying principles of the actions of a suicide bomber. However, there is a deeper problem, which is that trivial actions seem to have either no particular rational implications, or an infinity of rational implications. In other words, the problem of locating a principle does not go away just because we do not demand that that principle is part of the agent's psychology.

The second problem with the argument that maxims need not be psychologically manifested is even more serious. As we have noted, part of the importance of maxims is that they get us onto the Kantian conveyor belt: actions are actions on maxims; maxims are principles; principles require rational vindication and so on. However, if we are no longer saying that the maxim is something which the agent actively espouses, or which even figures in the agent's psychology at all, then the role of maxims in getting us onto the Kantian conveyor belt is undermined. If maxims are merely potential descriptions of action whose relevance is external to the agent's psychology, we can no longer see why the agent should be bound by any demand for their justification. This is especially the case when we consider that it is altogether uncertain what the maxim of certain actions could be, or whether such actions could be described as action on maxims at all. The impetus driving the Kantian conveyor belt is the demand for justification, and the reason that we get on it at all is that we have some proposition which requires justification. Yet, if we are accepting O'Neill's interpretation of Kant's theory then it is no longer clear that we have a proposition which requires justification. Although this interpretation still describes us as acting on a principle, what it means to act on a principle is considerably removed from our normal understanding of the phrase: all it can mean is that an act with certain underlying intentions has certain implications. But while it is possible to see how espousal of a principle, whether conscious or unconscious, incurs a demand and an obligation for some form of justification, it is rather harder to see that such demands and obligations are incurred by the implications of action. This does not mean that we escape responsibility for our actions and reasons. Those principles which we do consciously espouse, as well as those intentions we hold and those consequences we produce all invite demands for justification, but not necessarily the justification that Kant is after. A Kantian could respond to this argument by insisting that despite the lack of conscious psychological allegiance, to act in a way which has certain rational implications simply *is* to embody those implications, and to incur the obligation of justifying them. However, I think that at this point we are entitled to ask a question which is not just relevant to Kant, but to any argument which attempts to derive reasons from the supposed implications of

the very idea of action: why should we suppose that actions which carry no intention to embody principles should be treated as if they do embody principles?

So, it appears that there is at least one category of action which cannot be understood as action on a maxim, and which therefore escapes getting on the Kantian conveyor belt at all. It may seem that this is not enough to defend our account from the challenge presented by Kant's theory; after all, the internal reasons account is intended to account for all reasons; we do not argue that there are *some* actions which are not subject to external reasons, but that there are *no* external reasons. It may seem that we have left open the possibility that other categories of action are subject to external reasons. I agree that we have not shown categorically that all actions are capable of escaping the demand for justification associated with principles and their expression through action. However, I also think that we have shifted the burden of argument back to Kant and his defenders. Their starting position was that *all* actions were such that they were subject to the escalating need for justification represented by the Kantian conveyor belt, and we have shown that this is not the case. It remains for the Kantians to show that we should hold *any* actions to this standard of judgement. For, remember, the category of trivial actions was only selected as a test case because it was easiest to show that actions in this category could not be understood as actions on general principles. As we have seen in our discussion of Williams' example of the man who saves his wife from drowning, it is not hard to find situations which are practically and morally significant, yet in which it seems that to describe action as action on a principle is not only inappropriate but thoroughly misunderstands the situation. Furthermore, our reaction to such examples shows that actions which we do not think of as action on a principle are nevertheless subject to our judgement; that most of us immediately know what we think of the man saving his wife indicates that there are other dimensions of justification than rational justification of underlying principles, and that they often matter rather more to us. Sometimes action is just action, and we do not need to discover, or even look for, an underlying principle in order to judge whether the action is justified.

2.3 Motivations

All of the proponents of the types of theories we have considered so far strenuously deny or limit the role of motivations in producing reasons and action. Even Nagel who, as we saw, allowed that desires may always accompany action, saw those desires as a side-effect of deliberation, rather than as a determinant of reasons. By contrast, the final type of external reasons theory we will consider starts out by seeming friendly to the claim that motivations are necessary for the explanation of action and the existence of reasons, but goes on to argue for external reasons by supposing that our reasons depend on the motivations which we *should* have rather than the motivations which we *do* have, and furthermore, that the motivations we should have are fully determined by reason. In considering this type of theory we will concentrate on the work of its main current proponent, Michael Smith, and will be guided primarily by his recent book, *The Moral Problem* as well as some of the debate following its publication.[115] Smith is of particular interest to us because he articulates feelings and assumptions which I believe are shared by many philosophers but which are not often expressed explicitly. These intuitions are that the Humean account of motivational psychology is the only plausible one, but that at the same time there must be more room for objective normativity than Hume's account seems to allow; an apparent unease which corresponds to the tension between the individualistic and universalistic intuitions about reasons.

2.3.1 Smith's Argument for Ordered Motivations

We will have to go some way into Smith's theory to see the challenge it presents to our account of internal reasons. Our starting point is the clash which Smith sees as central to his argument: indeed, it is what he regards as the moral problem which gives his book its title. This clash occurs when we consider three propositions which express our intuitions about moral reasons, each of which may seem superficially plausible but which, when taken together, appear contradictory. Smith expresses these propositions thus:

1. Moral judgements of the form 'It is right that I ϕ' express a belief about an objective matter of fact, a fact about what it is right for me to do.
2. If someone judges that it is right that she ϕ then she is, *ceteris paribus*, motivated to ϕ.
3. An agent is motivated to act in a certain way just in case she has an appropriate desire and a means-end belief, where belief and desire are, in Hume's terms, distinct existences.[116]

In his response to *The Moral Problem*, 'The Metaethical Problem', Geoffrey Sayre-McCord helpfully identifies the first proposition as 'cognitivism,' the second

[115] In particular, see David O. Brink, 'Moral Motivation', David Copp, 'Belief, Reason and Motivation: Michael Smith's *The Moral Problem*', Geoffrey Sayre-McCord, 'The Metaethical Problem', and Smith's reply to all these papers, 'In Defense of *The Moral Problem*: A Reply to Brink, Copp, and Sayre-McCord'.

[116] *The Moral Problem*, page 12

proposition as 'internalism' and the third proposition as 'Humean psychology.'[117] According to Smith, the apparent problem with these propositions is that the third appears to be at odds with the other two. The third proposition makes motivation to act dependent on the existence of a desire, whereas the other two propositions imply that the judgement that an action is right can lead to motivation to perform that action without the involvement of any desires.

Smith attempts to resolve this apparent contradiction by following a line of thought which we have already considered; but he takes it further than we have been willing to take it. He acknowledges that there is an explanatory dimension to our talk of reasons as well as a normative dimension. However, he argues that the reason for our perception of these two different dimensions and the contradictions we encounter when we talk of them is that there are two distinct categories of reasons: motivating reasons and normative reasons. According to Smith, 'The distinctive feature of having a motivating reason to ϕ is that, in virtue of having such a reason, an agent is in a state that is explanatory of her ϕ-ing.'[118] By contrast, 'To say that someone has a normative reason to ϕ is to say that there is some normative requirement that she ϕ's, and is thus to say that her ϕ-ing is justified from the perspective of the normative system that generates that requirement.'[119] Smith stresses that when talking of normative reasons and motivating reasons he is not talking about different aspect of the same entities, but about entities of their own distinct kinds: 'whereas motivating reasons are psychological states, normative reasons are propositions of the general form, "A's ϕ–ing is desirable or required."'[120] Furthermore, he insists that there is no direct connection between normative reasons and motivating reasons: 'an agent may therefore have a motivating reason to ϕ without having any normative reason to ϕ, and she may have a normative reason to ϕ without having any motivating reason to ϕ.'[121] It is important to realise just how different Smith's understanding of reasons is from ours. Just as he does, we talk of reasons as possessing explanatory and justificatory dimensions, and, of course, we distinguish between different types of reasons; most notably between internal reasons and external reasons. However, we deny that external reasons exist, and insist that internal reasons carry the entire weight of explanation and justification, while Smith, in his division between motivating and normative reasons, has separated these aspects entirely. On his account, not only do internal and external reasons both exist, but they exist because they have different jobs to do.

Smith takes it that this separation of reasons into distinct categories solves the moral problem, because, if he is correct, the three propositions which stand in for our moral intuitions refer to different sorts of reasons, and therefore do not come into conflict. The intuition that reasons are objective refers to normative reasons. The intuition that desires are required to motivate action deals with motivating reasons. And the intuition that believed moral propositions motivate concerns the connection between the two. If we believe Smith, then it seems that the moral problem wasn't a problem after all. By making this distinction between normative and motivating reasons, Smith removes

---

[117] *The Metaethical Problem* in *Ethics* (October 1997), page 56
[118] *The Moral Problem*, page 96
[119] *The Moral Problem*, page 95
[120] *The Moral Problem*, page 96
[121] *The Moral Problem*, page 95

psychology from questions of justification and limits it to the realm of explanation. So, although we must address both types of reason in considering his theory, it is the possibility of normative reasons which are not dependent on psychology which presents the greatest challenge to our account; these are, of course, external reasons. We must explore Smith's theory a little further to see how he justifies the division of reasons into two categories, and how he accounts for normative reasons.

We will consider two of Smith's arguments for the distinction between normative reasons and motivating reasons, one of which draws on the familiar phenomenon of deliberative obstacles, and one of which depends on the claim that values and desires can come apart from one another. We have already acknowledged within our own account that deliberative obstacles present a difficulty to any straightforward account of practical reason, because they create an apparently contradictory situation in which we want to say that the agent has a reason to do something, but also that, because the reason is obscured by a deliberative obstacle, the agent may fail to take this reason into account but still be acting in an entirely rational and justified manner. Smith uses three examples to illustrate his understanding of this apparent problem. The first two of these examples are very similar to those we used in our own discussion: an agent wishes to buy a Picasso, but does not believe that the painting for sale is a Picasso, whereas it actually is; an agent wants a gin and tonic and believes that the gin bottle contains gin, whereas it actually contains petrol.[122] The man who wants a gin and tonic comes, of course, from Williams.[123] The third obstacle is rather different, as it implies a moral failing rather than a deliberative failing: an agent is standing on my foot but lacks any concern for my well-being and hence any inclination to get off.[124] In each of these cases Smith claims that our intuitions tell us that the agent has a reason to act one way (to buy the painting, to avoid drinking the petrol, and to get off my foot) but that we could not provide a rational explanation if they acted in accordance with those reasons. Despite our intuitions, their actions would be unintelligible. Smith's solution to the problem created by deliberative obstacles is to employ his two different categories of reasons. So, the agents in the examples have normative reasons to buy the painting, to avoid drinking the gin and to get off my foot. At the same time they have motivating reasons for turning down the painting, drinking the petrol and ignoring my suffering.[125]

The second argument of Smith's which we shall consider attempts to determine the nature of normative reasons. However, it resembles the first argument in that it deals with a phenomenon which seems to demand some sort of division within reasons if we are to account for it, and it explores this phenomenon and its implications through the use of examples. This argument attempts to draw a distinction between *values* and *desires*. Smith justifies this distinction by pointing to our common experience of *akrasia* (weakness of the will, or doing what we want to do rather than what we ought to do) and offers four examples borrowed from other writers: a kleptomaniac who steals even though he knows that it is wrong, a drug addict who hates his addiction but is enslaved by

---

[122] *The Moral Problem*, page 94.

[123] See 'Internal and external reasons', page 102.

[124] *The Moral Problem*, page 94.

[125] Of course, we have our own answer to this problem. In our account all reasons are of the same type, but some are obscured from agents by deliberative obstacles, while others are created by the existence of those same obstacles.

it, a woman who is seized by an urge to drown her baby, and a losing sportsman who is overwhelmed by a desire to attack his opponent.[126] These are all supposedly examples in which the agent desires that which he or she does not value. Smith also identifies the supposedly opposite phenomenon of valuing without desiring by citing the state of 'depression' in which, 'the depressive may know full well that the rational thing for to do is, for example, to get up and get on with her life: to go to work, to visit a friend, to read a book, to cook a meal or whatever,'[127] but is not motivated to do these things.

Smith draws three conclusions from the phenomena exhibited in these examples. First, he claims that values and desires are distinct. Secondly, he claims that values inform our normative reasons while desires inform our motivating reasons. And finally, and most surprisingly, he claims that values should not be regarded as motivational states at all, but rather as beliefs. This final conclusion is surprising because in our everyday talk valuing something seems to necessarily involve having a favourable disposition towards it: indeed, to deny that values have any motivational content seems at odds with our ordinary understanding of the term. However, Smith insists that the examples we have just considered, along with many others from our experience, show that desiring and valuing cannot be the same thing. Furthermore, if we accept the Humean account of motivation then we must also accept that the only entities which figure in motivated action are beliefs and desires, and therefore, if values are not desires then they must be beliefs. Smith acknowledges that there is a connection between value and desire, and claims that this connection is a rational one, saying that, 'Other things being equal, we rationally should desire what we value.'[128] In other words, if we intuitively feel that value has something to do with desire then we are right, because the possession of a value demands the possession of a desire. However, because this connection is a rational requirement it does not mean that a value is identical to a desire, or that a desire always accompanies a value. After all, any account of practical reason must allow that sometimes we are not rational.

So far this could be just another argument which attempts to show that external reasons exist by deriving such reasons from sources which are disconnected from motivations; and it would suffer from the same problems as those arguments because it would lack the explanatory dimension needed to connect it to the actual actions of agents. But, of course, this is exactly the sort of gap which Smith is concerned to bridge, and he attempts to do so in the next stage of his theory, in which he argues that normative reasons determine motivating reasons, and that values determine desires. This argument has six steps. We have already seen the first of these: the argument that values are beliefs rather than desires. And we have seen hints of the second step, which is to claim that, 'it is a platitude to say that what it is desirable to do is what we would desire to do if we were fully rational.'[129] The third step is to argue that, 'deliberation can produce new and destroy old underived desires.'[130] The fourth step is to claim that, 'by far the most important way in which we create new and destroy old underived desires is by trying to

---

[126] *The Moral Problem*, pages 133-134.
[127] *The Moral Problem*, page 135.
[128] *The Moral Problem*, page 148.
[129] *The Moral Problem*, page 150.
[130] *The Moral Problem*, page 95.

find out whether our desires are *systematically justifiable.*'[131] The fifth step is to argue that sets of desires are more systematically justified and rational to the degree to which they exhibit unity and coherence.[132] The sixth and final step is to argue that the unity of desires is a characteristic of full rationality and that a belief that a desire contributes to the unity and therefore rationality of the agent's set of desires is capable of producing that desire: 'For, if the analysis is right – an evaluative belief is simply a belief about what would be desired if we were fully rational, and the new desire is acquired precisely because it is believed to be required for us to be rational.'[133] So, in even more extreme summary, the argument seems to be that full rationality involves having a maximally unified and coherent set of desires, and that the ability of deliberation to both decide what this set of desires consists of, and to produce those desires in us, bridges the gap between the normative reasons provided by full rationality and the motivating reasons which produce action. As far as Smith is concerned, he has not only solved the moral problem by showing that the apparent contradiction in our intuitions is produced by confusion between normative and motivating reasons, he has also solved the problem of explanation by showing how we can act on normative reasons.

There are some aspects of Smith's argument which seem familiar both from ordinary experience and from our own account of internal reasons. We undoubtedly do critically assess our desires, and often wish that we desired differently from the way we do: the existence of self-help manuals, appetite suppressants and nicotine patches all attest to this phenomenon. Similarly, we acknowledge that our desires can be changed by deliberation, both through rational means (as when I discover my desire to be based on a false belief) and through less rational means (as when I realise that a proposed course of action resembles that taken by an admired figure, and this fires my enthusiasm for taking it). However, Smith's argument goes far beyond these familiar experiences and what can be allowed within our account. In some respects, Smith presents an argument which resembles that of the Kantian conveyor belt: if we allow that it is better to have some desires rather than others, we are immediately challenged to justify our preference for those desires, and then our preference for that preference, and so on, until we end up at the maximally coherent set of desires possessed by the fully rational agent. Unlike Kant, Smith does not offer a formula such as the Categorical Imperative which is capable of delivering substantive moral conclusions, but it is clear that his argument would not be unfriendly to such a concept. This possibility is further borne out, and the distance of Smith's argument from our everyday experience and our account of internal reasons further emphasised, when we realise that he is not talking about a maximally coherent set of desires peculiar to a particular agent, but is rather talking about a single coherent set of desires which should be possessed by all rational agents: 'we presuppose that fully rational agents would all have the same desires about what is to be done and desired in various circumstances.'[134]

It is at this point, where it is claimed that rational agents do not just have well ordered desires but that they have a common set of desires that it becomes most clear that, despite his apparent friendliness to the Humean account of motivation and action,

---

[131] *The Moral Problem*, page 95.
[132] See *The Moral Problem*, page 159.
[133] *The Moral Problem*, page 160.
[134] *The Moral Problem,* page 175.

Smith is an external reasons theorist. According to him, the normative reasons possessed by an agent are entirely independent of the agent's actual desires: 'the claim that an agent has a normative reason to ϕ is not a claim about her *actual* desires, but rather a claim about her *hypothetical* desires.'[135] It is, of course, possible within our account of internal reasons that an agent could achieve these hypothetical desires through sound deliberation from his or her starting motivations. But Smith goes further than this: he supposes that the agent possesses the normative reasons associated with the common set of desires, even if the agent could *never* apprehend those reasons or acquire those desires, no matter how soundly he or she deliberated. In Williams' terms, Smith insists that there are true statements about the existence of reasons for an agent which cannot be falsified by the presence or absence of elements of the agent's subjective motivational set.

The concept of the fully rational agent with the maximally coherent set of desires is particularly important because it is required not just by Smith's argument, but by any argument which attempts to derive external reasons from the rational ordering of motivations. We may, and shall, dispute some of the detailed steps which Smith takes towards this concept, but even if we show his argument to be flawed, the thought may remain that motivations should be ordered in some way which is not itself dependent on motivations. If such ordering truly is independent of the contingent motivations of individual agents then it will inevitably converge on some form of ideal which, if it recognises variations between individuals at all, at most does so on the basis of circumstances. The fully rational agent with maximally coherent desires may live on, even if not in the context of Smith's argument.

Like all of the arguments we have discussed so far, Smith's argument, and the more general position which it exemplifies, presents a serious challenge to our account of internal reasons because, if correct, such arguments overcome the problem of explanation and the disconnection between reasons, motivations and action which trouble other external reasons theories. However, I believe that Smith's argument is not only flawed in its own right but also that it reveals general problems with arguments of this type, and our response will consequently start by challenging two key distinctions within Smith's argument, before moving on to consider what is wrong with the more general idea that our motivations could or should be fully determined by reason.

2.3.2 Challenging the Division Between Value and Desire; Between Normativity and Motivation

As we have seen, Smith divides value from desire, by arguing that values are

[135] *The Moral Problem*, page 165. We may note with surprise that Smith presents this principle as one implied by Williams in 'Internal and external reasons', albeit at a point when he is taking issue with Williams' argument. However, I believe that Smith has misinterpreted Williams here. He has taken Williams' allowance that deliberation may produce changes in the agent's motivational set, and therefore that reasons are not derived directly from desires, to be compatible with his view that agents can reach a maximally coherent set of desires and full rationality through deliberation. This understanding goes wrong in two ways: it fails to acknowledge that, for Williams, the changes produced in the agent's motivational set are still a function of starting motivations; and, more importantly, that Williams' emphasis on indeterminacy means that he would be unlikely to accept the concept of full rationality as a determinate state, and certainly not one that involved the possession of a uniform set of desires.

beliefs, and divides normativity from explanation, by arguing that there are two distinct categories of reasons. These divisions are important to us because they challenge the central claim of the internal reasons account: that there are no external reasons. We naturally suppose that our values are sources of reasons, and if values are beliefs with no necessary connection to the contents of our subjective motivational sets, then reasons may exist which are not dependent on those contents. And the point of Smith's distinction between normative reasons and motivating reasons is to show that reasons can exist independently of our motivations. However, we can show that we do not need either of these divisions.

Smith's argument for the existence of two distinct categories of reason on the revolved around three examples in which agents were separated from reasons which we would intuitively ascribe to them in ways resembling the deliberative obstacles we discussed earlier at some length: the art buyer, the gin drinker, and the inconsiderate person who won't get off my foot. This argument also depended on the apparent confusion in our intuitions about the reasons of the agents in these examples. However, as we have already seen through the concept of weakly internal and strongly internal reasons, we can explain why apparently contradictory reasons exist while preserving the connection between explanation and normativity, without needing to invoke distinct categories of reasons. It is important to remember that we are not just arguing about terminology here: strongly internal reasons do not equate to motivating reasons as Smith understands them, nor do weakly internal reasons equate to normative reasons. The point about the distinction we have made is that both types of reasons have both an explanatory and a normative dimension, but are distinguished by their degree of accessibility to the agent. The difference between our distinction and Smith's distinction shows that Smith misses the significance of the confusion caused by his examples. The confusion we experience when we confront these examples does not arise because we find an explanation or a justification which are at odds with one another: if that was the case then we might not be confused at all. We are confused because we find reasons which have both explanatory and normative aspects, yet which clash, and we do not suppose that we can resolve this confusion simply by pulling explanation and justification apart. If Smith is to be believed then there is no normative dimension to the reasons of the art buyer who walks away from a painting which he does not believe to be a Picasso; all we can offer is an explanation. But from the perspective of his ignorance the art buyer's action is perfectly justifiable: he chooses not to buy a painting which he does not want. To say that the art buyer's reason not to buy the painting is not at all normative is to omit an essential part of our understanding of the reason, and consequently our intuitions about it.

The person who won't get off my foot is rather more persistently troublesome, as he is evidently not subject to any deliberative obstacle. If he cared about my pain then he would have no difficulty in working out that he ought to get off my foot. The problem is, of course, that he doesn't care. So, we cannot account for his actions or our confusion about the reasons he has in terms of strongly and weakly internal reasons. However, this does not mean that we cannot account for him at all. Indeed, we have two ways of doing so. Firstly, if we allow that reasons are dependent on motivations, and therefore that not everyone has the same reasons, we can see that our intuitions about what people have reason to do are most likely based on our common experience of what people commonly have reason to do, on the basis of a common, but not inevitable, set of human

motivations. So, most people do care about the pain of others, at least to the extent of not causing it obviously and gratuitously, so would have reason to get off my foot. Even those few who didn't care about my pain would care about the disapproval causing pain invites, so would still have a reason to get off my foot. So, if we find our intuitions about the reasons possessed by the agent confusing in this case, it may simply be that our intuitions don't work in the case of such an odd agent: someone who neither cares about the pain of others nor about the approval of society largely escapes our intuitions (but, we hope, not our institutions). The second way of dealing with this unusual agent is to point out that even though our intuitions may be misleading about what he has reason to do, our intuition may be entirely sound in leading us to say that he ought to get off my foot, but that this intuition does not have to be expressed in terms of reasons. In other words, we do not have to say that the agent is failing to follow a valid reason to say that there is something wrong with his failure to get off my foot. We are not limited to judgements of reasons, but can say that he is callous, selfish, or even cruel. Taken together, these ways of accounting for our judgements about the man who won't get off my foot mean that we cannot straightforwardly suppose that the specific intuition that he has a reason to get off my foot is a reliable indicator that such a reason actually exists; there are other possibilities which might confound our intuitions.

So, it seems that, while we can acknowledge that we have some puzzling intuitions when presented with examples such as those considered by Smith, we can solve these puzzles perfectly well using the existing materials of our account of internal reasons: we do not have to do anything so extravagant as to invent separate categories of reasons. Our challenge to Smith's second division, between desires and values, takes a similar approach; we can ask whether Smith's examples of valuing without desiring and desiring without valuing really have the consequences he supposes them to. That is, we can see whether we can account for the examples he uses without having to suppose that values and desires are entirely different orders of entity. The obvious alternative to Smith's argument is to suppose that values are a more settled form of desires, and that we are better off describing them both as species of motivation rather than using the language of desire. After all, we pointed out right at the beginning of our discussion that it is more useful to speak of motivations than of desires. However sincere we are in attempting to use 'desire' as a general, technical term to distinguish a dispositional set of mental entities from propositional entities such as beliefs, it still carries connotations of urgency, immediacy and even weakness, to the extent that it can lead us to conclude that any mental entity which does not share these characteristics must be a belief. Once we see values and desires as differing patterns of motivation then we are able to deal with Smith's examples quite easily. In those cases where we desire something we do not value, it simply means that we have an immediate urge which lacks support from any more settled or persistent disposition. Sometimes this may be regrettable, especially when the urge conflicts with our more settled dispositions, but this does not mean that values and desires come apart to the extent that we must suppose that they are different entities.

However, just as with the man who wouldn't get off my foot, there is a rather more persistently troubling case: that of the state of depression, in which we apparently value without desiring. It is tempting to say that sometimes our more settled dispositions simply do not manifest themselves as immediate desires, but that would be too glib. If someone supposedly possessed a persistent value which never manifested itself in any

immediate impetus to action, even when the circumstances so coincided with the value as to appear to warrant that action, we would rightly wonder whether the agent actually held the value at all. However, this very thought indicates the way in which we should deal with this example: we should ask whether it represents a conceivable situation at all. Remember, Smith is attempting to show that values and desires come apart to the extent that we can only make sense of them as distinct types of entities. This means that, for the depressive to genuinely hold the value in the absence of desire, he or she must do so without that value producing any trace of motivation. As we have seen, Smith imagines that, 'The depressive may know full well what the rational thing for her to do is, for example, to get up and get on with her life: to go to work, to visit a friend, to read a book, to cook a meal, or whatever. But the effect of her depression may be precisely to remove any desire at all that she has to do any of those things.'[136] Smith needs that 'at all' to close down the question of whether values could be a form of desires, but it is that 'at all' which renders his argument implausible. For we are not just being asked to imagine a depressive state in which the agent is experiencing some sort of affective cloud which masks her desires, or even that her desires are weakened beyond the point where they can overcome the most basic level of inertia. We are being asked to imagine that the agent has no desires *at all* which could be satisfied by going to work, cooking a meal and so on. Indeed, rather than constituting a recognisable situation which any theory of motivation and reason must account for, this example seems to constitute a bizarre artefact of Smith's theory. It seems far more plausible to say that the agent is in an unfortunate motivational state in which the affective impact of depression puts the more immediate desires that would move her to action out of joint with the more settled motivations that constitute her values.

In passing, we should also note that Smith's attempt to split values from desires and normativity from motivation undermines his original intention to reconcile the Humean account of motivation and action with an account of objective reasons. Although Smith claims to have achieved this, I do not think that he acknowledges quite how far he has departed from Hume's theory. The Humean aspect of his theory has been reduced to a mechanistic explanation of how desires produce action, with this mechanism being dominated by a rational ordering of desires. Of course, it is not a legitimate criticism on its own to say that Smith has departed from the work of a writer he claims to endorse, and which we happen to endorse as well. We should be guided by the truth rather than by blind faithfulness to a particular writer or a particular theory. However, in departing from Hume, Smith has abandoned some of the most attractive aspects of his theories, including the possibility of explanation which drew him to Hume in the first place. In particular, Smith has accepted the thoroughly anti-Humean claim that beliefs alone can produce desires. It may seem odd to describe this claim as anti-Humean; after all, Hume's account allows that passions can be destroyed by reason, when reason detects that those passions are founded on false beliefs, and our own Humean account of internal reasons allows that motivations can be modified by deliberation. However, neither we nor Hume allow that such effects can be produced by reason *alone*; within our account, the impact of deliberation on motivations occurs because it transmits the influence of motivations which already exist. So, if I follow food fads in the popular press, I may change my attitude towards drinking a glass of wine a day from approval to disapproval, but this is

[136] *The Moral Problem*, page 135.

only because I care about my health, or losing weight, or whatever effect drinking or abstaining is supposed to produce. If I don't care about these things then my attitude is not affected. By adopting a position in which desires are produced by deliberation purely on the basis of values, where values are understood to be beliefs without any motivational content, Smith simply moves the mystery of explanation one step further out. We may be able to explain actions by reference to desires, but we are still puzzled about how the beliefs which Smith supposes to constitute values could bring these desires about.

2.3.3 Challenging the Demand for Motivations to be Rationally Ordered

It is important to show that the distinctions between normative and motivating reasons and between values and desires on which Smith builds much of his argument are mistaken as, if they were correct, they would have serious consequences for our account of internal reasons. However, we cannot take ourselves to have disposed of Smith's argument and arguments of its type simply by dealing with those distinctions; the main challenge to our account of internal reasons arising from such arguments is the claim that motivations can and should be ordered by reason, and this claim could be made on grounds other than those used by Smith. We shall deal with this challenge by pointing out the consequences of supposing that our motivations are as subject to rational ordering as Smith, or anyone who wishes to derive external reasons from rational constraints on motivations, must suppose them to be, and by showing that these consequences are at odds with our understanding and experience of our reasons, our motivations and even our identities. For it is important to realise how strong the implications of the position which Smith represents are; such positions do not simply allow, as Hume does, that under certain restricted circumstances our motivations may be unreasonable, or, as allowed by Williams and our internal reasons account, that deliberation may influence and direct motivations. Rather, they imply that motivations should be *fully determined* by reason. The extremity of such positions is illustrated by the figure of the fully rational agent with the maximally ordered set of motivations. The significance of this figure is that, if we were fully rational, we would have motivations like those of the fully rational agent; that is, we would all have the same set of desires. And, as well as clashing with our earlier thought that the contingency and indeterminacy of our reasons is, in large part, constitutive of our freedom, this idea contradicts our experience and understanding of motivations and reasons in three main ways.

Firstly, it implies that irrationality is endemic. According to Smith we are rational insofar as we possess and act on a set of desires which can be systematically justified; the same set of desires as that possessed by the fully rational agent. According to Smith's account, fully rational agents are those agents which have maximally coherent desires and, as we have seen, he argues that there is only one such set of desires. Fully rational agents will therefore all have the same desires. Yet the diversity of human desires is a familiar feature of everyday life; although we expect that many desires are shared, and that many objects of judgement will generate motivational consensus, we also expect that, at least in matters of detail, desires will differ from agent to agent. I do not expect, for example, that anyone should feel the same way about my family as I do, not just because they are not in the same circumstances as me, but because they are unique individuals who would feel differently even in the same circumstances. More to the point,

because they are unique individuals, I do not even expect that they will feel the same way about their family as I do about my family. This means that, on Smith's account, we are less than fully rational; and, considering the wide variations in desires which exist, a very long way from fully rational. This is not logically impossible: we could theoretically all be irrational, just as we could all theoretically be moral failures. Indeed, it would not be implausible to claim that we all routinely exhibit some symptoms of irrationality to greater or lesser degrees some of the time. However, I think that our common understanding of rationality is that most of us are mostly rational most of the time, and, indeed, that the complex and ordered societies in which we live would not exist as they do[137] if this were not the case. Smith's theory does not imply that we experience occasional lapses of rationality but rather that, if we persist in desiring in all our various and chaotic ways, we are constitutionally irrational. Furthermore, we not only expect that people have different desires, but regard it as a positive aspect of human existence; for many of us, the idea that we desire differently is an expression of the values of diversity and individuality, and the idea that we should not desire differently stands in opposition to these values. We may express these values in many different ways, including passionate defences of individualism, lofty ambitions for tolerance, or even through clichéd sayings such as, 'It takes all sorts to make a world.' The point is that we value difference as well as expecting it, to the extent that the prospect of a world in which everyone shared the same set of desires sounds more like satire or bad science fiction than an image of rational harmony.

The value that we place on the difference between individuals brings us to the second problem with Smith's claim that we should all be like the fully rational agent with the maximally coherent set of desires: as well as contradicting our experience and values concerning others, it violates our understanding of our own identities. Smith argues that the contingent desires which we happen to have are incidental to the desires which we ought to have, and that those contingent desires should therefore be considered part of our circumstances rather than part of us: he says that, 'what we have reason to do is relevant to our circumstances, where our circumstances may include aspects of our own psychology.'[138] In an example considering why one agent might have reason to visit a wine bar rather than a pub, he also says that, 'the crucial point in this case is that a relevant feature of your circumstances is your preference for wine, whereas a relevant feature of my circumstances is my preference for beer. That this is a relevant feature of our circumstances is manifest from the fact that I can quite happily agree with you that if I were in your circumstances – if I preferred wine to beer – then the fact that the local wine bar sells very good wine would constitute a reason for me to go there as well, just as it constitutes a reason for you.'[139] Of course, it is possible to play the game of pretending to have the desires of others; but only up to a point. For those desires which do not matter very much to me, it is indeed possible for me to imagine what I would be like if those desires were different, or if I did not possess them at all. But we do not have to go very

---

[137] Of course, this qualification that our complex, ordered societies would not exist *as they do* is an important one. Complex and ordered social entities can exist without rationality: beehives and ant-hills provide two obvious examples. But these societies do not exist as human societies do: the members of such entities are closer to mechanical components than active, rational participants.

[138] *The Moral Problem*, pages 170.

[139] *The Moral Problem*, pages 170-171.

far to discover those desires which matter a very great deal to me, and which, if I lost or modified them, would constitute a significant upheaval in my identity. If someone else imagines what it would be like to possess those desires, it seems most natural to say that he or she is not imagining what it would be like to be in my circumstances, but is rather imagining what it would be like to be *me*. It is worth noting in passing that this tendency to regard our desires as somehow incidental to our identities is common to the external reasons theorists we have considered, from Kant's treatment of desires as 'alien causes'[140] to Korsgaard's claim that, 'Hume has no resources for distinguishing the activity of the person *herself* from the operation of belief, desires, and other forces *in her*.'[141] In Hume's account and in our account of internal reasons we do not make this distinction because we recognise, in accordance with our everyday experience, that desires constitute part of the agent's identity.

The final problem with the argument that we should all possess the maximally coherent and systematically justified set of desires to be considered rational, is that we routinely judge ourselves and others to be rational while accepting the existence of desires within an agent which contradict one another, and which would therefore be considered less than fully coherent or systematically justified on Smith's account. In our ordinary use of the term, the mere possession of contradictory desires by an agent is not enough to justify a judgement that the agent is irrational. Indeed, we may consider that it is an essential part of deliberation to find a way to the practical conclusions which are best capable of satisfying a contradictory set of desires. For example, consider the prudent saver who wishes to ensure that he has enough money for his retirement and who finds the prospect of investments which combine a high rate of risk with a high rate of return both frightening and attractive. Or consider the writer who yearns to see the piece that she is working on finally finished, yet hates the process of writing and is continually tempted by distractions. Or, finally, simply consider the mixture of fear and delight with which someone contemplates a rollercoaster ride. The point of these examples is not just that the agents experience a mix of contrasting desires; it is that there is no obvious way in which deliberation could reach a conclusion which could quell one or the other of these desires, and, furthermore, we do not typically demand or expect such a conclusion from deliberation. The saver cannot eliminate the risk associated with his investment, so must take a judgement which accommodates his conflicting desires as much as it anticipates the future. The writer knows that she has to overcome her temptation to procrastinate and get on with her work, but this does not mean that the temptation is irrational or that it goes away. And to suggest that the rollercoaster rider should reach a choice between fear and anticipation is to miss the point of rollercoasters. It could be claimed that Smith's argument could accommodate such examples. After all, it concerns the requirement for coherence rather than examples of what counts as coherence. So, it might be that the fear and anticipation felt in the queue for a rollercoaster are part of a maximally coherent set of desires, which includes the desire for having the best possible time on a rollercoaster. However, it is difficult to see that this allowance could be pushed too far before the requirement for maximal coherence becomes meaningless. One suspects that Smith's fully rational agent would not find him or herself in the queue for a rollercoaster at all.

At this point we should note that, in the course of defending his position from a

---

[140] *Groundwork*, 4:446.

[141] 'The Normativity of Instrumental Reason' in *Ethics and Practical Reason*, page 233.

variety of challenges, Smith has claimed that the degree of uniformity in the desires of fully rational agents demanded by his theory is not as great as it might initially appear. He says that, 'the convergence required is very circumscribed. There is no suggestion that fully rational people will all have the same tastes in food, and clothes and basketball teams. On the contrary, they will presumably be at least as culturally and individually diverse as human beings throughout history have been.'[142] It is not clear whether Smith has modified his view of the desires of the fully rational agent in response to challenges, or whether he is simply making explicit claims which were implicit in the earlier argument. However, whatever the reason for these further claims, I do not think that they make the claim that we should be like the fully rational agent with the maximally unified and coherent set of desires any more plausible, attractive or recognisable. Smith, of course, cannot allow that we have just any variations in our desires; to do so would be to undermine his entire argument. So, he allows that fully rational agents can possess varying desires, 'only if fully rational creatures regard it as permissible for people to have and to act on such desires, that is, only if they are at least indifferent to people having such desires, and in favour of their acting on them once they have them.'[143] In other words, within the desires of the fully rational agent there are some which must be common to all such agents, but there are others which can be allowed to vary without disrupting full rationality.

The problem with this view is that it attempts to deal with the inescapable fact that agents which we would normally judge rational nevertheless exhibit wide variations in their desires by marginalizing those desires; variant but rational desires are placed in a category which is not significant to the overall unity and coherence of the agent's desires and which therefore, it is implied, can safely be considered part of the agent's circumstances rather than part of his or her identity. Smith says, 'Characterise a choice situation in its entirety –'What would we desire ourselves to do in a situation in which the external circumstances are thus and such (list them completely) and we have these and those desires and beliefs and other mental states (list them completely)?' – and, I say, fully rational creatures will all converge on a desire that the very same course of action be pursued.'[144] However, those desires in which we most conspicuously vary are not equivalent to our tastes in food, clothes and basketball teams, and cannot be so easily marginalized or treated as part of our circumstances. Those desires which most characterise us as individuals are those manifested in our relationships with other people; not just our likes, but our loves and loyalties. Such strong desires and the undeniable variations in their manifestation in different agents do not fit Smith's model in either its original or its revised version; Smith seems committed to either denying that such desires matter, when they plainly do, or insisting that they must be uniform in all rational agents, when they plainly aren't. And it seems inevitable that anyone who follows Smith by arguing that our desires should be fully determined by reason will end up in such a position; at some point any such theory must attempt to mark out and defend a set of motivations which we must all possess to be considered rational, and such an assumption of uniformity will always be subject to the charge that it is not what we demand or expect of rational agents, unless the set of motivations upon which they insist is so minimal as to

[142] 'In Defense of The Moral Problem', page 89.
[143] 'In Defense of The Moral Problem', page 89.
[144] 'In Defense of The Moral Problem', page 89.

be of no consequence.

2.4 Summary: Answering the Challenge from Reason

As we said at the outset of this part of our discussion, we have not considered all possible external reasons theories. However, I hope that we have not only shown that our internal reasons account can resist challenges from a particularly important set of external reasons theories, but have also indicated some general problems with this group of theories. Two general problems correspond to the common intuitions about reasons we introduced at the beginning of our discussion: the individualistic intuition and the universalistic intuition.

The first general problem within the theories we have discussed is that they seem to have no prospect of satisfying the individualistic intuition. As we have seen, at some point each of these theories supposes that the same reasons apply to all of us, and that the possession of particular reasons by particular agents is determined by the circumstances in which they find themselves rather than who they are. Furthermore, in order to deal with the unavoidable variation we experience in the reasons which we ascribe to ourselves and to others, they tend to treat aspects of our personalities and lives which we regard as fundamental to our identities as part of our circumstances, thus contradicting not only our individualistic intuition about reasons, but also our informal intuitions about the nature of human identity. The implication of such theories is not only that external reasons exist in the sense that they are defined in our internal reasons account, but that all reasons are to some extent external to us, in the sense that they are determined independently of who we are, and that it is our duty to find out which of them apply to us and to act accordingly. If these theories are correct then agents are not unique individuals with reasons which reflect that uniqueness but interchangeable units who could be slotted into the circumstances of one another's lives without noticing. The individualistic intuition about reasons indicates that we think we are more than that, and that anyone who insists we are not would need to offer rather more justification than we have seen in the work of our external reasons theorists.

The second general problem is that the external reasons theorists we have considered seem to be subject to a particularly strong version of the universalistic intuition; much stronger than we find in our everyday experience of reasons, to the extent that their theories all come to rest on a particular unsubstantiated assumption: that reasons must be universal. This assumption leads to theories such as those we have seen, in which it is supposed that the way to find out how we should act is to find the foundations which underpin universal reasons, whether those are inescapable aspects of identity, our autonomous rational natures, or maximally coherent set of motivations, and to proceed to discover or test individual reasons for actions in the context of these foundations. Yet, despite our universalistic intuition about reasons, we have been given no justification for supposing that reason must be universal to this degree, or that our actions require such foundations. We must remember that the universalistic intuition is only an intuition that our reasons, despite their proximity to our individual nature and circumstances, are nevertheless subject to some form of external judgement. The assumption that all reasons must be universal to the extent that they rest on foundations which transcend our individual nature goes so far beyond this intuition that the reasons it leads to and the various means by which those reasons are discovered and justified are barely recognisable from the perspective of our ordinary understanding of reasons.

However, we must also remember that the universalistic intuition still remains only partially satisfied by our account of internal reasons. Fortunately, although we have rejected each of the external reasons theories we have considered, they are still of some value to us in solving this problem. They remain eloquent, if extreme, expressions of the universalistic intuition, and I believe that an attempt to diagnose why people hold external reasons theories and go to so much trouble to defend them, as well as why people make what sound like external reason statements in everyday talk, will help us to understand the common motivations underlying the universalistic intuition, and to see how they could be satisfied within the terms of our account.

3. Being Reasonable

3.1 What We Want from Reasons

So, we will now attempt to diagnose the motivations, attitudes and assumptions which underlie the universalistic intuition, both as it is expressed in the work of external reasons theorists, and as it is expressed in everyday talk of reasons. In doing so we are, of course, operating entirely within the bounds of our account of internal reasons. That account insists that reasons are dependent on the contents of our subjective motivational sets, so an understanding of the motivations of those people who argue for external reasons theories or who make apparent external reasons statements should help us to understand their reasons for doing what they do; assuming, of course, that they are acting for reasons. I also hope that this understanding of what we want out of reasons when we give voice to the universalistic intuition will give us some idea of how to better satisfy this intuition within out internal reasons account.

3.1.1 What External Reasons Theorists Want from Reasons

The most obvious attitude towards reasons within the work of our external reasons theorists is an aversion to contingency and arbitrariness. It is present throughout their theories as they quest for objectivity and the universal foundations of reasons, and is explicitly expressed by Kant when he says that, 'Everyone must grant that a law, if it is to hold morally, that is, as a ground of obligation, must carry with it absolute necessity,'[145] Korsgaard expresses a similar thought when she argues from the claim that 'Unless something attaches normativity to our ends, there can be no requirement to take the means to them,'[146] to the conclusion that 'There must be unconditional reasons for having certain ends, and, it seems, unconditional principles from which those ends are derived.'[147] Smith explicitly identifies arbitrariness as anathema to normativity: 'the only decisive point we can make about normativity is that arbitrariness, as such, always undermines normativity.'[148] In comparing his own position to that of Kant at the start of *The Possibility of Altruism*, Nagel notes that although his theory and Kant's theory contain different claims about the self-conceptions which we necessarily hold, they are united by the idea that these self-conceptions are necessary. He goes on to say that, 'different as they are, both are thought to be conceptions which we cannot escape and are thought to provide that basis for ethical motivation which in other internalist theories is provided by various motives and desires. Because of the alleged inescapability of these conceptions, a view of the Kantian type entails that we are not really free to be amoral, or insusceptible to moral claims. That is what makes us men.'[149] As we have seen, the aversion to contingency and arbitrariness underlies the work of external reasons theorists to the extent that they do not set out in their enquiries to discover the nature of reasons and their origins whatever they may be; they set out with the hope and expectation that

---

[145] *Groundwork*, 4:389.
[146] 'The Normativity of Instrumental Reason', page 251.
[147] 'The Normativity of Instrumental Reason', page 252.
[148] 'In Defense of The Moral Problem', page 90.
[149] *The Possibility of Altruism*, page 14.

reasons must be objective and universal, and that therefore their origins must lie in the unconditioned and the inescapable. Where our writers look for the basis of reasons tells us as much about their underlying motivations as the ways in which they look. Once again, this is evident from Nagel: 'It will in any case not do to rest the motivational influence of ethical considerations on fortuitous or escapable inclinations. Their hold on us must be deep, and it must be essentially tied to the ethical principles themselves, and to the conditions of their truth. The alternative is to abandon the objectivity of ethics.'[150]

That somewhat unfortunately phrased claim in one of our quotations from Nagel ('That is what makes us men.') also expresses the second desire concerning reasons which we can detect within the work of external reasons theorists. It is strongly related to the aversion to contingency, but is rather directed towards one of the sources of contingency within Humean theories of action, and within our account of internal reasons: motivations. There has long been a tension within philosophy between theories which give primacy to motivations and those which give primacy to reason,[151] with proponents of the latter type of theory, including, of course, our external reasons theorists, evidently fearful that giving any ground to motivations other than that required for the mechanical explanation of action is to surrender our rational nature at best to grubby selfishness, and at worst to mere animalistic reflex: these seem to be the alternatives if we do not heed the call of 'what makes us men.' In our discussion of their theories we concentrated on the positive arguments of our theorists for sources of reasons which are independent of desires, but they often spend nearly as much time attempting to show that desires could not only never be adequate sources of reasons, but that they get in the way of the reasons which we do have. We have already seen an example of this in Smith's argument for the ways in which values and beliefs come apart, and can see it again in examples of obstructive desires used by Korsgaard: 'You want to see the movie but you are too idle to go into town; you want to go out with him but you are too shy to call and ask him for a date; you want to work but depression holds you in its smothering embrace.'[152] While we can see that idleness, shyness and depression are all regrettable, Korsgaard wants to go further than this: she wants to show that these particular motivations fail to give us reasons, that they obstruct legitimate reasons and, furthermore, that they show us how unsuitable motivations are for determining reasons at all. The desire for reasons to rise above the supposedly tawdry concerns of motivation is further shown by Nagel who, when arguing for altruism, takes his opponent to be egoism, which apparently, 'holds that each individual's reasons for acting and possible motivations for acting, must arise from his own interests and desires, however those interests may be defined.'[153] If those interests are understood broadly enough then this is a crude description of our own position, as it is of anybody whose argument is roughly Humean: but few of these positions could be properly described as egoistical. For Nagel, it seems that egoism is all that you can end up with once you allow that desires can produce reasons.

The third motivation which we can detect in the work of external reasons theorists

---

[150] *The Possibility of Altruism*, page 6.

[151] For a lively and recent discussion of this conflict, see the section 'Dionysus and Apollo' in Simon Blackburn's *Ruling Passions*, especially pages 88-89.

[152] 'The Normativity of Instrumental Reason', page 229.

[153] *The Possibility of Altruism*, page 84.

is the hope that, as reasons are not contingent or dependent on anything as base as motivations, they constitute inescapable moral obligations. This hope is most clearly expressed by Korsgaard when she says that, 'Reason has the power to compel obedience, and to punish us for disobedience. It in turn is bound to govern us by laws that are good. Together these facts yield the conclusion that the relation of the thinking self to the acting self is the relation of legitimate authority.'[154] This hope also has a counterpart: the fear that if reasons are not obligations then anything goes. We are all somehow free from moral and rational judgements, or worse, we are left without any direction whatsoever. Korsgaard again expresses this fear when she considers whether someone who did not accept any rational normative authority would therefore be obliged by consistency to commit suicide, and concludes that, 'There is nothing the normative sceptic *has* to do. But it is worth remembering what an extreme position practical normative scepticism is. The normative sceptic has no reason for doing anything.'[155] It is dangerous to ascribe specific desires on the evidence of theory, and the external reasons theorists could claim that in discovering the obligations imposed by reasons they are simply following the course of their enquiries. But there is an evident relief in their writing, especially that of Korsgaard, when they reach the conclusion that our lives are governed by reasons which have the force of law; not only are we unable to escape our obligations, but we can go on living after all.

The fourth set of motivations apparent in the work of external reasons theorists could be described as a desire for intelligibility, but a special sort of intelligibility that goes beyond our normal understanding of the word. We might normally expect that reasons are intelligible when we can understand them though a sufficient acquaintance with the nature and circumstances of the agent. However, external reasons theorists want an intelligibility that does not rest on such contingent aspects of the world. This is most clearly expressed in the work of Kant. It is no accident that, when making his case for the autonomy of rational beings, he claims that each agent, 'has two standpoints from which he can regard himself and cognise laws for the use of his powers and consequently for all his actions; *first*, insofar as he belongs to the world of sense, under laws of nature (heteronomy); *second*, as belonging to the intelligible world, under laws which, being independent of nature, are not empirical but grounded merely in reason.'[156] In other words, for Kant, if reasons rest on those aspects of the world to which we would normally look for explanation but which have no further rational justification, then they are not, in his terms, intelligible. We can always ask why a contingent aspect of the world provides a reason, and will never find an answer that satisfies Kant. Of course, we have shown that we do not need to even begin to ask Kant's question, and that we can ordinarily consider reasons to be justified and intelligible on the basis of considerations for which we neither have nor require further justification. However, this observation is likely to do little to quell the desire for a particularly uncompromising form of intelligibility within Kant and his followers.

The final set of motivations found within the work of our external reasons theorists may seem rather less transcendent and rather more pragmatic; it is the hope that reflection can come to an end and issue in conclusions. However, just as with the desire

---

[154] *The Sources of Normativity*, page 165.
[155] *The Sources of Normativity*, page 163.
[156] *Groundwork*, 4:454.

for intelligibility, this hope seems to take a particular form for external reasons theorists; they do not want deliberation to reach conclusions which are merely satisfactory, they want deliberation to reach conclusions which are *perfect*, in the sense that they are the only possible conclusions which could be reached if the agent was deliberating correctly. The expression of this hope sometimes seems to be tinged with a little desperation; some of the external reasons theories we have considered allow reason to operate unchecked in its most destructive mode. The hope for conclusions is the hope that once unfettered reason has swept away all other possible foundations, there will be something left. The tension between the consequences of radical reflection and the need for conclusions is expressed by Korsgaard when she says:

> I desire and I find myself with a powerful impulse to act. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse doesn't dominate me and I have a problem. Shall I act? Is this desire really a *reason* to act? The reflective mind cannot settle for perception and desire, not just as such. It needs a reason. Otherwise, as long as it reflects, it cannot commit itself or go forward.[157]

In other words, we must reflect, otherwise we do not have a reason, but that reflection must reach a conclusion, otherwise we do not have a reason either. This need to find some basis for action in the face of radical reflection is further expressed by Korsgaard when she says that moral scepticism, 'is the view that the problems which reflection sets for us are insoluble, that the questions to which it gives rise have no answers. It is the worry that nothing will count as reflective success, and so the work of reflection will never be done. It is the fear that we cannot find what Kant called "the unconditioned."[158] Korsgaard here takes herself to be diagnosing the condition of the sceptic about practical reason, but I think that she is also coming as close as anybody does to explicitly expressing one of the powerful combinations of belief and desires which drives external reasons theorists: the thought that reason can undermine all those things we ordinarily think of as justifications for action, and the hope that we can nevertheless find foundations somewhere.

Furthermore, the writers we have considered seem to want not only that deliberation can come to a conclusion, but that the process of coming to this conclusion can produce agreement among agents. This hope is most explicitly articulated by Michael Smith when he sets out what he takes to be the platitudes of objectivity:

> 'When A says that ϕ-ing is right, and B says that ϕ-ing is not right, then at most one of A and B is correct'; 'Whether or not ϕ-ing is right can be discovered by engaging in rational argument'; 'Provided A and B are open-minded and thinking clearly, an argument between A and B about the rightness or wrongness of ϕ-ing should result in A and B coming to some agreement on the matter'; 'The rightness of someone's ϕ-ing is determined by the circumstances in which that person acts, circumstances that might be faced by another.'[159]

---

[157] *The Sources of Normativity*, page 93.
[158] *The Sources of Normativity*, page 94.
[159] *The Moral Problem*, page 39-40.

Smith's concern in *The Moral Problem* is to reconcile the apparent contradictions in our common intuitions about morality, and by the end of the book he takes himself to have achieved this: 'in conjunction with some plausible assumptions about the potential that moral argument has to bring about agreement, the analysis also allows us to think that our moral talk is in fact legitimate. For it is plausible to suppose that through moral argument we can in fact discover what the reasons that we all share really are.'[160] There is evidently a desire here for moral talk to be legitimate, and for this legitimacy to be expressed by the potential for deliberation, in this case public argument, to produce agreement. Of course, for external reasons theorists the agreement produced by shared deliberation should be a particular type of agreement: not arbitrary consensus but the joint recognition of reasons that exist independently of agents and their deliberations.

Kant expresses a similar desire when he claims that, 'There is no one – not even the most hardened scoundrel, if only he is otherwise accustomed to use reason – who, when one sets before him examples of honesty of purpose, of steadfastness in following good maxims, of sympathy and general benevolence (even combined with great sacrifices of advantage and comfort), does not wish that he might also be so disposed,'[161] We must remind ourselves once again that we are not directly concerned with the truth of Kant's claims but with his underlying motivations. However, we should note that this is one of Kant's claims which is most clearly false: even if we had not argued for the existence of the sensible knave it is simply implausible that examples of honesty, steadfastness, sympathy and benevolence will produce yearnings in the heart of every hardened scoundrel. His espousal of this obviously false claim is therefore a powerful illustration of his hope for the transforming power of reason. We should also note that Kant's theory does not even rely on reason having this power; it would be entirely compatible with his theory for all of us to be scoundrels who ignore the call of reason. Yet he continues to hope; a hope that reaches its culmination in the ideal image of a Kingdom of Ends which is far from any kingdom we are likely to encounter in this world.

In some ways, the motivations, attitudes and assumptions regarding reasons we have found within the work of external reasons theorists resemble those which we might find in anybody concerned with finding reasons for action: an aversion to arbitrariness; a hope that reasoning will produce conclusions and agreement; and a desire for reasons and deliberation which are intelligible. However, as we have seen, what is given voice in the arguments of external reason theorists is rather different from that which we might expect to find in our everyday talk of reasons. We can characterise this difference by observing that all of the attitudes, assumptions and motivations concerning reasons displayed by external reasons theorists shows that they want above all out of reasons is that they are *absolute*; at the risk of overstating the case, they want reasons which are immune to historical contingency, which are inscribed on eternity, in much the same way that some religious people believe that the words of holy books such as the Qu'ran and the Torah existed before they were transcribed by humans: prior, unchanging and inescapable. They want to be assured that in any given situation for any agent there is a correct course of action, that the agent cannot avoid the reason to pursue this course of action, and that if the agent could only apprehend and follow this reason then he or she would be guaranteed of acting correctly and with justification. Such certainty dispels arbitrariness

[160] *The Moral Problem*, page 202.
[161] *Groundwork*, 4:454.

because it ensures that reasons are only derived from those considerations capable of providing absolute justification; it guarantees intelligibility not necessarily because individual agents can follow deliberative paths to the correct reasons, but because correct deliberation necessarily leads to these reasons, even if it is beyond the power of a particular agent to follow; and it produces conclusions and agreement because, once the correct reasons have been apprehended and accepted, there is nothing more to say without courting irrationality. But, despite our universalistic intuition, I do not believe that the rest of us want or seek this level of certainty from our reasons. To see this, let us explore the motivations, assumptions and attitudes we express in our everyday talk of reasons, particularly when we say things that sound like external reasons statements.

3.1.2 What We Usually Want from Reasons

In our everyday talk of reasons we say things such as, 'Of course you've got a reason to do it,' or, 'This is the only rational thing to do,' or, 'Don't be stupid, you've got no reason to act like that.' Even though in making such statements we do not make any formal claims about the dependence of reasons on motivations, we often pay little attention to the agent's motivations in making them; our talk often implies that assertion of the existence of a reason should persuade agents and silence objections.[162] We are making statements that sound like external reasons statements. Of course, within our account of internal reasons we allow both that such statements may be bluff and that agents often *should* be persuaded by assertions and demonstrations of reasons because those assertions and demonstrations clear away deliberative obstacles to reveal reasons that were previously obscure. However, we must also acknowledge that people making such statement often want more out of them than the elimination of deliberative obstacles: they want, in accordance with the universalistic intuition, that the subjects of their statements should fall in line with their judgements about reasons. We must attempt to determine whether the universalistic intuition has the same force in people making such statements as it apparently does in external reasons theories by asking whether they are underpinned by the same motivations, attitudes and assumptions.

It may seem difficult to find motivations such as an aversion to contingency and arbitrariness within everyday talk of reasons, as the nature of reasons is not a topic which we discuss every day. However, such evidence can be found, especially if we make life easy for ourselves by considering exceptional cases such as novel moral dilemmas. Novel moral dilemmas are those which we have not encountered before, and in which, although we feel that we ought to have strong, definite views, we are left somewhat at sea. Examples include questions such as how the reservations of wealthy citizens of the industrialised world regarding genetically modified crops should influence the potential use of this technology to feed millions of people in absolute poverty. It is in grappling with this sort of question that we feel the desire for a secure basis for reasons most

[162] Of course, these are not the only things we say about reasons. In accordance with the individualistic intuition we expect that the right reasons for an agent will be specific to him or her. When asked to offer advice about reasons, we often try to find out what it is that the agent wants. However, we are primarily concerned here with the everyday statements we make that approximate to external reasons statements and which express the universalistic intuition, not those that approximate to internal reasons statements and express the individualistic intuition.

strongly: our thoughts and feelings regarding such matters are often barely formed, to the extent that we are not quite sure what we think, or whether the various arguments that we tentatively offer are really engaging with the problem at all. So, I may be confident in my views about the ownership of genetic patents and the exploitation of farmers, while still being entirely uncertain of how to compare the risks of genetic modification against the lives of those who could be helped; or even how to regard the whole practice of genetic modification. In such circumstances I may desire a secure basis for reasons not just to settle the question but to ease my embarrassment in not even knowing what my view is. I certainly do not feel that such questions can be settled by invoking arbitrary considerations: any considerations that I bring to bear in deliberating about such questions need to be justified themselves. Consequently, at least in such circumstances, I exhibit an aversion to arbitrariness.

However, this aversion is not quite the same as that which we found within the work of external reasons theorists. That aversion led to the pursuit of that which is not only not arbitrary, but also that which is not contingent in any way: the unconditioned. By contrast, our everyday aversion to arbitrariness in deliberation and the discovery of reasons is simply that: a rejection of those considerations which appear to have no bearing on the matter at hand. This rejection of what is purely arbitrary does not prevent us from accepting that which is merely contingent, such as the influence of motivations. Indeed, if we consider novel moral dilemmas once again, we can see that part of the problem they present, and of the source of our unease when we face them, is that we don't know what we *feel* about them as well as not knowing what we *believe* about them. Even in more mundane situations we can readily see how we not only allow but expect that factors peculiar to individuals determine reasons. If someone asks our advice on questions such as where to go on holiday, or what sort of job to look for, or where to buy a house, we are likely to ask questions about his or her preferences and circumstances, and to tailor our advice about what to do - the statements we make about the reasons we think he or she has for action - accordingly. Few, if any, of us would feel the temptation to over-ride such contingent considerations by seeking the unconditioned. We would not attempt to identify the perfect holiday or the perfect house for all agents in those circumstances, or consider whether going on holiday or buying a house was the best course of action for the agent at all: we would advise that agent on the basis of our understanding of him or her.

We find a similar situation when we consider our attitudes towards the relationship between motivations and reasons. Within the work of our external reasons theorists we found a suspicion of motivations, to the extent that if allowed a role at all this was restricted to an instrumental function in the production of action. It may appear that we can find a similar suspicion of motivations within our everyday talk of reasons. We may think that somebody who claims to be acting charitably is actually acting out of self-interest, or we may think that strong motivations such as anger or hatred inhibit someone's ability to think coherently. However, these suspicions are, of course, different from those entertained by external reasons theorists. These suspicions do not imply that reasons cannot be rooted in motivations, but rather that reasons, deliberation and statements about reasons may be obscured or distorted by motivations. We will typically allow that the agent who claims to be charitable but is actually being selfish has reasons to be selfish; what we are suspicious of is what the agent says about his or her reasons.

The same motivations that give the agent reasons to act selfishly give the agent reasons to lie about his or her reasons. Similarly, we do not suspect that the agent's anger is not a source of reasons, but rather that the experience of anger may be so strong that it prevents the agent from deliberating soundly to his or her best reasons for action, even in respect of that anger. So, we have cause to be suspicious of motivations within our everyday understanding of reasons, but this does not stop us from acknowledging those motivations as sources of reasons.

The pattern we have found so far is repeated when we consider the desire for intelligibility, in that we find just such a desire expressed within our everyday talk of reasons, but one that has a significantly different emphasis from that found within the work of our external reasons theorists. The desire for intelligibility is clearly apparent within our everyday talk of reasons, as much of this talk is expended tying to gain an understanding of the deliberations and reasons of others. Oft-repeated questions such as, 'Why did you do that?' or, 'Why do you think you should do that?' are typically intended to draw out the lines of deliberation that led to a particular conclusion about action, in the hope that this will make the agent's supposed reasons intelligible to us.[163] Furthermore, such attempts to achieve intelligibility are not confined to our judgements of the reasons of others; we often question our own half-formed and unarticulated deliberation and reasons to make them more explicit to ourselves. However, in attempting to find intelligible reasons, we do not suppose that these reasons will only be intelligible if everything concerned with their production is itself governed by reason. Rather, we find reasons that come to an end in contingent aspects of the world such as motivations entirely intelligible. Indeed, it is the identification of just such a basis in motivations that we *need* to make the actions and reasons of others intelligible to us; the question, 'Why did you do that?' is typically intended to discover what the agent wanted as much as how the agent deliberated. The chances of understanding reasons and actions are particularly improved if we share or even just sympathise with the motivations of another agent. However, we can find reasons intelligible even if we cannot achieve such sympathy. We can see how the motivations of mass murderers and terrorists lead to their terrible conclusions about how they should act, even if those motivations and the deliberation which follows from them horrify us. Such difficult cases also remind us that although we desire intelligibility, we do not expect the reasons of others to be transparently intelligible to us at first sight. On the contrary, we expect that sometimes intelligibility may only be achieved through discussion and debate, particularly when we are dealing with unfamiliar circumstances and cultures. We only really fail to find intelligibility when deliberation has gone dramatically wrong, or when motivations lie so far outside our sympathies that we cannot grasp them at all; and in this latter case, we may acknowledge that it is the agent's motivations that are unintelligible rather than the reasons that follow from those motivations, and even then that their apparent unintelligibility may be due to our failure to understand rather than because they could not be understood by anyone. So, while we want intelligibility out of reasons, it is the intelligibility provided by the ability of agents to understand one another and themselves, rather than the intelligibility supposedly

[163]Of course, we should note that such questioning often not only fails to render action intelligible, but also fails to draw out the line of deliberation actually followed by the agent; challenges to actions and reasons often produce post-hoc rationalisations which bear little resemblance to why the agent actually acted.

provided by transcendental rational principles.

The pattern continues when we consider the hope that reasoning can issue in conclusions and that reasoned conclusions can produce agreement. This hope is explicitly manifested in our everyday talk of reasons. Even though we occasionally engage in debate for the pleasure of the argument, and even though such knowingly fruitless debate can sometimes be the most passionate, most of the time our public and private deliberation would simply make no sense if we did not hope and expect that it would issue in conclusions. The difference between the hopes and expectations of everyday talk and those expressed in external reasons theories is that for external reasons theorists the conclusion of deliberation and agreement on action is a product of discovering the only right reasons to follow in the situation, while on our everyday understanding of action agreement is reached and deliberation concluded simply because we have found reasons which those people participating in deliberation can accept as good enough, even if there are other potential conclusions which could be reached, and which could also be accepted as good enough. This does not mean that our reasons are not serious or that they are compromised in some sense; the questions that prompt our deliberation may be very serious indeed, and we may steadfastly defend those reasons on which we have settled. The essential point is that even the most passionate conviction about our reasons will not be held because it is supported by the sort of theoretical justification demanded by our external reasons theorists; such passionate convictions will be held partly because of the agent's confidence in the deliberation which produced them, and partly because they concern things that the agent cares about.

So, it appears that, although in our everyday talk of reasons we express motivations, attitudes and assumptions which resemble those expressed in the work of external reasons theorists, we give them a significantly different emphasis: we have an aversion to arbitrariness, but we are comfortable with difference and contingency; we are suspicious of motivations when they are not sincerely expressed or when they disrupt reasoning, but we are not suspicious of them simply because they are motivations; we want reasons to be intelligible, but find that an understanding of character, situation and motivations makes reasons more intelligible than the application of transcendental rational principles; and we want deliberation to reach conclusions and we want those conclusions to be right, but we can be confident in those conclusions without needing to show that they are necessarily true for all people at all times. In short, the motivations, attitudes and assumptions we express in our everyday talk of reasons seem to afford reason something resembling the instrumental role allowed by Hume. Above all, we want reason to be practically effective, and our motivations seem to express an awareness, conscious or unconscious, that the burden we place on reasons must be delicately judged; we must demand of our reasons that they are sufficiently robust to fit the situation, but must not demand so much that we undermine our confidence without warrant.

I believe that we can best characterise the motivations, attitudes and assumptions regarding reasons which are expressed in everyday talk, by saying that what we ordinarily want out of reasons is that they are *sustainable*, and that this sustainability stands in contrast to the absoluteness which external reasons theorists seemed to want from reasons. Sustainable reasons are those which agents can accept as sufficiently justified to suit the circumstances under consideration and, furthermore, which are capable of standing up to those challenges which might be expected to be mounted

against those reasons. Each of the motivations, attitudes and assumptions we have found in our everyday understanding of reasons contributes to the achievement of sustainable reasons. So, the avoidance of arbitrariness helps us find reasons which cannot readily be undermined, as they do not have an arbitrary basis. Similarly, the sincere expression of motivations in deliberation and the expression of reasons means that those reasons are not susceptible to challenge by the revelation that they depend on hidden motivations. We can better sustain those reasons which are intelligible and which are based on intelligible deliberation because we can demonstrate and understand how those reasons were come by. And those reasons which constitute practical conclusions and bring us to agreements are sustainable because they give us what we want out of reasons.

The challenges which sustainable reasons stand up to need not be explicit or made by people other than the agent who is deliberating; it is a familiar part of deliberation that we challenge ourselves by asking whether our deliberation is really sound or our conclusions are really justified. Sustainability does not guarantee that reasons are capable of withstanding all conceivable challenges; just those which seem appropriate to the situation. So, if I am choosing a birthday present for a friend, then my reasons are sustainable if I can justify to myself or any other interested person why I have chosen a particular gift, or spent a certain amount of money, and they remain sustainable even though I could not justify the institution of celebrating birthdays by giving gifts. However, a previously sustainable reason may be undermined by an unexpected yet warranted challenge, and thereby come to be unsustainable. If, when shopping for a birthday present it was pointed out to me that I was about to spend a substantial amount of money on a momentarily amusing piece of trivial rubbish, when I might do something which is perhaps more boring and worthy, but was ultimately more useful and compassionate, such as making a contribution to charity in my friend's name, I might find that my previously sustainable reason no longer stood up to scrutiny.

This means that, unlike the immoveable and unchanging edifice of reasons imagined by external reasons theorists, those reasons which can be considered sustainable because they conform to our common motivations, attitudes and assumptions concerning reasons vary from agent to agent and even within individual agents over time. But this is what we expect to find within any mature individual facing complex and shifting practical situations, and with an understanding of the world which is continuously developing; it reflects our experience of a rational life rather better than the idea that all our reasons aare laid out for us and that it is merely our duty to discover them. Moreover, the thought that reasons can shift between and within individuals yet are nevertheless guided by our desire for sustainability provides us with a way in which reasons can be closely aligned to the nature and circumstances of individual agents, yet be subject to external judgement and normative pressure; in other words, to offer a way of satisfying both the individualistic and universalistic intuitions while remaining compatible with the internal reasons account. In order to develop this thought further we shall adopt a theoretical framework which not only provides us with a vocabulary to describe and assess our common patterns of motivation and behaviour, but which also returns us to the roots of our account: the theories of David Hume and, in particular, his account of virtue.

3.2 Virtue

It may seem a large and unexpected step from the consideration of the patterns of motivation regarding reasons to the overtly moral concept of virtue. However, as we shall see, if we continue to follow Hume, common patterns of motivations and our attitudes towards them are exactly what constitute virtue and vice. Furthermore, recent philosophical enquiries have questioned the emphasis placed on morality in discussions of virtue and have asked whether it is correct to perpetuate the distinction drawn by Aristotle between moral and intellectual virtues,[164] or whether a more comprehensive and integrated understanding of virtue is required.[165] The existence of such recent work on intellectual virtues reflects the increasing attention paid to virtue in recent decades,[166] which makes it particularly important to be clear about the understanding of virtue we are using in our discussion. So, we will start by summarising the Humean understanding of virtue, before going on to consider whether the patterns of motivations we have discussed and the behaviour they produce constitute a virtue in Humean terms. I believe that they do constitute a virtue and, unsurprisingly, this claim raises some questions which we shall also attempt to answer: whether we can justify calling what we have found a virtue rather than a skill; whether we have found a natural virtue or an artificial one; whether by introducing the concept of virtue we have unwittingly introduced another source of external reasons; whether the virtue fits our account of internal reasons and our intuitions about reason; and, finally, how we should judge our external reason theorists with respect to this virtue.

3.2.1 Humean Virtues

Hume's account of virtue is one of the optimistic aspects of his theory, in which he builds an understanding of complex human behaviour from basic elements, as opposed to the pessimistic aspects of his theory, in which the rational basis for fundamental elements of our experience such as causality and existence is torn down, tempting us to doubt and despair. It has its foundation in Hume's most fundamental moral claim: that morality is not rooted in reason but in the passions, as expressed in the title of the very first section of the book of the *Treatise* which deals with morals, *Moral distinctions not deriv'd from reason.*[167] This claim is unpalatable to many, and gives rise to similar reactions to those which we have already seen in relation to the Humean account of action. However, just as we have found with this account, we do not have to accept that the stark claim lying at its centre is the whole of the theory. We have seen how a sophisticated account of action, reason and motivation can be built upon the central claim that reason is subordinate to subjective motivation in the arena of action, and a similarly sophisticated account of morals can be built upon the claim that moral judgements are also dependent on subjective motivations. Fortunately, unlike our account of action, in

---

[164] *Ethics*, 1103a14-b1, page 91.

[165] For example, see Linda Zagzebski's book *Virtues of the Mind*. From time to time throughout this discussion we will consider the implications of this recent work for our own enquiry, and shall use Zagzebski's book as a representative of thinking in this area.

[166] The beginning of this renewed interest in theories of virtue is usually traced to G.E.M. Anscombe's paper 'Modern Moral Philosophy', published in 1958.

[167] *Treatise*, Book III, Part I, Section I, page 507.

which we had to do much of the construction on top of the central claim ourselves, in the case of morals Hume has done much of the building for us. Large parts of the *Treatise* and most of the second *Enquiry* are concerned not just with setting out Hume's theories of virtue, but with employing those theories to identify and explain the virtues we recognise. We do not have to accept all of Hume's claims about particular virtues to see the merit in his theory, and this theory gives us a model for defining the virtue concerned with practical reason.

We can summarise Hume's account of virtue by considering five essential elements. The first of these is the basic claim we have already briefly encountered: that moral judgements are expressions of passion rather than of reason. This claim has led Hume's theory to be associated with various labels such as 'emotivist' or 'expressivist',[168] although such labels are often unhelpful as they are more properly applied to subsequent theories, some of which are no longer recognisably Humean. Hume offers two arguments for the claim. The first is familiar from our account, as it concerns the relationship between reason and action. Hume argues that moral judgements are essentially practical: in his terms they are, 'supposed to influence our passions and action and to go beyond the weak and indolent judgements of the understanding,'[169] and by this stage in his theory he has already established that reason alone is not capable of giving rise to action. To invoke another potentially unhelpful label, we could say that Hume is expressing a form of internalism about moral motivation; according to him, for something to count as a moral judgement by an agent it must contain some sort of impetus capable of inclining that agent towards action, even if the agent never acts in accordance with it, and as reason alone is incapable of providing that impetus it can be ruled out as the source of moral judgements. Hume's second argument is based on his theory of the understanding rather than his theory of the passions, and depends on observations of what it is possible for us to perceive in situations that call for moral judgements. He claims that, 'If the thought and understanding were alone capable of fixing the boundaries of right and wrong, the character of virtuous and vicious must either lie in the relation of objects, or must be a matter of fact which is determined by our reasoning,'[170] and then goes on to argue that if we deploy our reasoning to discover such relations or matters of fact we simply cannot find them. He says, 'Take any action allowed to be vicious. Wilful murder, for instance. Examine it in all lights and see if you can find that matter of fact, or real existence, which we call vice. In which-ever way you take it you find only certain passions, motives, volitions and thoughts. There is no other matter of fact in the case.'[171] This passage continues in a way which shows that Hume believes that moral judgements do exist, even though they cannot be derived from reason alone, and shows where he believes them to be derived from: 'The vice entirely escapes you, as long as you consider its object. You can never find it, till you turn your reflection into your own heart and, and find a sentiment of disapprobation, which arises in you, towards this action.'[172] In other words, moral judgements are determined by feelings rather than by the conclusions of reason.

---

[168] For example, see Philippa Foot in *Natural Goodness*, pages 5-6.
[169] *Treatise*, Book III, Part I, Section I, page 509.
[170] *Treatise*, Book III, Part I, Section I, page 515.
[171] *Treatise*, Book III, Part I, Section I, page 520.
[172] *Treatise*, Book III, Part I, Section I, page 520.

The second element of Hume's account of virtue is the claim that those judgements which are distinctly moral are directed towards the *character* of agents and actions rather than towards the individual agents and actions themselves. ''Tis only when a character is considered in general, without reference to our particular interest that it causes such a feeling or sentiment, as denominates it morally good or evil.'[173] So, if I disapprove of the personal loss I suffer when my pocket is picked I am not making a moral judgement, whereas if I disapprove of the act as an example of robbery, or of the person who robbed me as a thief, I am making a moral judgement. In practice, of course, I am likely to make all of these judgements together. This element of the theory rids us of any suspicion that moral judgements based on motivations must necessarily be expressions of self-interest: our motivations are sophisticated enough to include attitudes of approval and disapproval towards general characteristics of actions and agents which may have no direct bearing on our interests.

The third element of Hume's account of virtue is not explicitly articulated, but is rather implied by the list of virtues and vices which he attempts to account for: that virtues and vices are sufficiently recognisable patterns of motivation and behaviour for us to be able to give names to them. So, in the *Treatise*, Hume deals with, among others, justice, injustice, loyalty and chastity. And his approach to the exploration of many of these virtues and vices is one which is open to misinterpretation: the telling of stories, intended to be illustrative rather than factual, about how we might have come to regard these patterns of motivation and behaviour as virtues and vices. It is partly in the construction of such stories that Hume presents the fourth element of his account, and divides virtues and vices into those which are *natural* and those which are *artificial*. Natural virtues and vices are those which will be valued or despised in all human societies, while artificial virtues and vices are produced by contingent circumstances. So, for example, Hume imagines an initial 'state of nature'[174] in which humans were bereft of the advantages of civilised society but not of the capabilities and basic inclinations which could produce such society, and goes on from this state to describe the origins of justice and property. That Hume does not mean this story to be taken literally is indicated when he says that, 'philosophers may, if they please, extend their reasoning to the suppos'd state of nature; provided they allow it to be a mere philosophical fiction, which never had, and never cou'd have any reality.'[175]

Perhaps the best examples of Hume's division of the virtues into the natural and the artificial are the virtues of benevolence and justice, of which Hume claims the former to be natural and the latter, in a move which he obviously expects to be controversial, to be artificial. Benevolence is taken to be a natural virtue partly because it is universally prized, to the extent that, 'even its weaknesses are virtuous and amiable,'[176] but also because it is taken to underpin and regulate other virtues: 'A propensity to the tender passions makes a man agreeable and useful in all the parts of life; and gives a just direction to all his other qualities, which otherwise may become prejudicial to society.'[177] The claim that justice is an artificial virtue may seem less controversial when we realise

---

[173] *Treatise*, Book III, Part I, Section I, page 524.
[174] *Treatise*, Book III, Part II, Section II, page 544.
[175] *Treatise*, Book III, Part II, Section II, page 544.
[176] *Treatise*, Book III, Part III, Section II, page 655.
[177] *Treatise*, Book III, Part III, Section II, page 653-654.

that Hume is not using the term in our modern sense, with its connotations of fairness, but to refer to the conventions and behaviour regarding the preservation and respect of private property. Nevertheless, he evidently expected a similar reaction to that which he would likely get today to his categorisation of justice as an artificial virtue, as he felt compelled to defend it as well as to argue for it.[178] Of course, the change in attitude over time towards the concept of justice lends weight to its categorisation as artificial: Hume's 18th century contemporaries would doubtless be inclined to regard our understanding of justice as artificial, just as we would theirs. Hume's argument was slightly different to this, though: it was that it is theoretically possible to imagine a golden age, albeit a fictitious one, in which 'every man had a tender regard for another,' and, 'nature supplied abundantly all our wants and desires,'[179] and consequently the conventions and inclinations which establish private property would not be needed, and justice as he understood it would not be regarded as a virtue.

So, we may readily agree that justice as Hume originally conceived it is an artificial virtue. It is worth pausing here to consider Hume's defence of this categorisation, however, as it serves to dispel some illusions about the use of the term 'artificial,' and, by extension, some concerns about Hume's whole scheme of virtues. There are three such illusions. Firstly, the artificiality of some virtues could be interpreted as somehow meaning that they are less significant than the natural virtues, or that moral judgement influenced by these virtues is somehow less legitimate. However, Hume's claim that moral judgements are produced by sentiments does not distinguish between sentiments according to their origin. It does not matter whether the feeling of approbation or disapprobation has its ultimate origins in human nature or inescapable yet contingent characteristics of human circumstances; all that matters is the feeling. Hume himself is concerned to avoid the misinterpretation: 'To avoid giving offence, I must here observe that when I deny justice to be a natural virtue, I make use of the word *natural* only as oppos'd to *artificial*. In another sense of the word; as no principle in the human mind is more natural than a sense of virtue, so no virtue is more natural than justice. Mankind is an inventive species; and where an invention is so obvious and absolutely necessary, it may as properly be said to be natural as any thing that proceeds immediately from original principles, without the intervention of thought or reflection.'[180]

The second illusion is to suppose that because the existence of a virtue such as justice is explained by reference to the more efficient satisfaction of the interests of the members of a just society, that judgements about justice are actually judgements about self-interest. Hume's scheme requires not that the sentiments which constitute moral judgements are masks or proxies for the satisfaction of more basic interests, but that they are actual feelings about what they seem to be about. Again, Hume is at pains to point this out: 'This self-interest is the original motive to the establishment of justice but a sympathy with the public interest is the source of the moral approbation that attends that virtue.'[181] In other words, the explanation of the origin of a virtue may refer to entirely different motives from those which promote behaviour in accordance with the virtue once it is established.

---

[178] See *Treatise*, Book III, Part II, Section II, page 545.
[179] *Treatise*, Book III, Part II, Section II, page 545.
[180] *Treatise*, Book III, Part II, Section I, page 536
[181] *Treatise*, Book III, Part II, Section II, page 551.

The third illusion is to imagine that because, as Hume does, we can attempt to trace the lineage of artificial virtues, and because such attempts could reasonably be called processes of rational investigation, the artificial virtues could be acquired by reason. However, there is an enormous difference between understanding how and why a particular society came to value, applaud and exhibit a particular virtue and coming to acquire that virtue oneself. Acquisition of the motivations which underpin artificial virtues may require long habituation and immersion in a culture, especially as those motivations associated with the virtue may be far removed from those which lie at its origins. Hume points this out in his account of the development of justice, in which the passion of self-interest both leads to and gives way to the passion of sympathy with public interest.[182] The idea that we can understand virtue without acquiring it, and possibly without even being able to acquire it, is the Humean version of a phenomenon we could label the tragedy of virtue. Whatever particular theory of virtue we are adopting, it is common to acknowledge that virtues are achievements of character, and that they are largely acquired by means such as our upbringing and the lessons of our earliest experiences. This means that for most people, by the time we have reached a sufficient level of maturity to grasp the concept of virtue, it is too late for us to genuinely acquire those virtues which we do not already possess.

The final element in Hume's account of virtue has already been anticipated in the recognition of natural virtues, and the thought that even artificial virtues may be recognised in most conceivable human situations: that these virtues and the motivations which underpin them allow the formation of a *common point of view*. Elements of the common point of view appear in the *Treatise*, but it is most clearly described in the second *Enquiry*, when Hume says that in making a moral judgement an agent must, 'depart from his private and particular situations and must choose a point of view common to him with others; he must move some universal principle of the human frame, and touch a string to which all mankind have an accord and sympathy.'[183] Part of the idea of the common point of view as expressed here reflects the idea mentioned earlier, that sentiments expressed in moral judgements are directed towards the character of acts and agents rather than the acts and agents themselves. However, the argument for the common point of view goes beyond this thought; it claims that someone making moral judgements, 'expresses sentiments, in which he expects all his audience to concur with him.'[184] Note that Hume is not claiming that there is a single, definite, common point of view from which moral judgements can be made, and which supersedes judgements made from the point of view of individual agents. Rather, the idea of the common point of view expresses Hume's optimistic belief that human nature is roughly uniform; or at least sufficiently uniform to provide a consensus in matters which are as important to us as the virtues. The substantive components of the common point of view do not have to be (and possibly could not be) fully defined, and we are encouraged to remember that the strings to which all mankind have an accord and sympathy include the artificial virtues. The scope of the common point of view is shown in the *Dialogue* accompanying the second *Enquiry* in which Hume presents caricatures of ancient Greek and (for him) modern French society. He does this in order to point out the differences between them and the

---

[182] See *Treatise*, Book III, Part II, Section II, page 547 and 551.
[183] *Enquiry Concerning Morals*, Section IX, Part I, page 272
[184] *Enquiry Concerning Morals*, Section IX, Part I, page 272

judgements they would make of each other, but also to point out that a large part of their differences can be explained by material and social circumstances, and to further point out that despite their differences they share fundamental characteristics that provide the basis for their own understanding of virtue.

We have now seen enough of Hume's account of virtue to realise that, if we can apply it to our common motivations concerning reason, we have a way of satisfying the universalistic intuition. When we judge the reasons of others, and expect them to respond to our judgements we may, in part, be making judgements about their achievement and expression of a virtue; of patterns of motivation and behaviour of which we generally approve. Before we apply this account to the motivations we have found concerning reasons, though, we must do two things. Firstly, we must ask whether what we have seen of the account gives us grounds for accepting it. The most contentious part of the account is the initial claim that moral judgements are sentimental. Although this claim is undoubtedly compatible with our account of internal reasons, it is not the same as the question of whether motivations are required for reasons to exist, it is highly controversial, and to attempt to do it justice would take us far from our central topic of reasons. Fortunately, we do not need to settle the claim to engage with the rest of Hume's account of virtue. For this account is not like the Kantian conveyor belt in which a line of reasoning leads inexorably to a conclusion but is disrupted if we challenge any of the steps along the way. Rather, most of Hume's account of virtue is based on the naturalistic observation that we exhibit common patterns of motivation and behaviour, and that there are common evaluations of those patterns. The claim about the nature of moral judgement influences our understanding of what is going on when we make those evaluations, but does not affect the question of whether we actually make them. And I think that it is clear that we do make those evaluations. When we think about justice or honesty, or any other virtue, we think first about their human manifestation in patterns of motivations and behaviour rather than the general concepts. So, although the claim about the sentimentality of moral judgements is important for Hume's overall account, we can to some extent separate it from his account of virtue: all we need is the thought that virtue is a general pattern of motivation and behaviour.

The second thing we must do before we attempt to apply the Humean understanding of virtue to our common motivations concerning reasons is to acknowledge that most of the discussion of virtue in recent years has been conducted in Aristotelian rather than Humean terms, and, furthermore, that at least some of this discussion is friendly to the proposition that external reasons exists. We must consider, therefore, whether by introducing the concept of virtue, we have introduced yet another source of external reasons.

3.2.2 Hume and Aristotle

There are many similarities between Hume's account of virtue and Aristotle's. They both regard virtue as the language of moral judgements and the vehicle of the moral worthiness of agents, as well as something which is embedded within character rather than expressed through individual actions. They also both take as the raw material of their enquiries an understanding of virtue which is not abstract, but which is embedded in the societies in which they live. However, there are two important differences. Firstly,

Aristotle has a teleological understanding of virtue. For Hume, the explanation of virtue and the judgements we make in relation to virtue start with our passions and develop according to our histories and social circumstances, with the result that we cannot expect that virtues must necessarily drive towards a particular end; while harmonious passions might be easier to satisfy, and the consequent virtues easier to attain, there is no requirement for the nature of history and humanity to be that tidy, or for things to turn out that way. By contrast, Aristotle starts from the thought that virtues are virtues to the extent that they produce and constitute the proper end for human agents. According to Aristotle, the proper end for human agents is the condition of *eudaimonea*, a term which is notoriously hard to translate, but which in contemporary virtue ethics seems to be most frequently rendered as 'flourishing.'[185] Eudaimonea is partly constituted by a desirable state of psychology and character, in which actions, reasons and motivations are aligned with virtue. There is something of this idea in Hume. Towards the end of his discussion of morals in the *Enquiries*, Hume considers the rewards of virtue:

> But in all ingenuous natures, the antipathy to treachery and roguery is too strong to be counterbalanced by any views of profit or pecuniary advantage. Inward peace of mind, consciousness or integrity, a satisfactory review of our own conduct; these are circumstances, very requisite to happiness, and will be cherished and cultivated by every honest man, who feels the importance of them.[186]

However, that last phrase: 'every honest man, *who feels the importance of them*,' indicates the difference between Hume's position and Aristotle's. For Hume virtue and the peace of mind that may accompany its attainment are a fortuitous, if predictable, consequence of our typical passions and circumstances, rather than an inescapable product of our natures. Furthermore, while the desirability of the condition of eudaimonea may lead us to seek it, for Aristotle it is rather more than just desirable: it is the proper end for human beings as determined by our nature as a species, and the appropriateness of this end to us carries its own normative weight. For Aristotle, we should not pursue eudomainea because we want to be that way; we should pursue it because it is the way that we should be.

The second difference between Hume and Aristotle is that, whereas Hume's account of virtue is founded in the passions and only allows a limited role for reason, Aristotle understands virtue as dependent on a particular form of rationality: *phronesis*, or the virtue of practical wisdom, which constitutes the ability to recognise the wise and virtuous course of action called for in a particular situation. It allows an agent, 'to be able to deliberate rightly about what is good and advantageous for himself; not in particular respects, e.g. what is good for health or physical strength, but what is conducive to the good life generally.'[187] This aspect of Aristotle's theory leads to what is probably the greatest difference between the work of modern followers of Aristotle and Humean theories such as the one we have developed here. Such theories, especially those of John McDowell, are cognitivist in relation to morality, meaning that they regard moral considerations as facts which can be grasped and known through the correct perception of

---

[185] For example, see Rosalind Hursthouse in *On Virtue Ethics*, pages 9-10.
[186] *Enquiry Concerning Morals*, Section IX, Part II, 233, page 283.
[187] *Ethics*, page 209.

the pertinent aspects of morally significant situations or choices. Furthermore, as these judgements are made through the exercise of phronesis, which is an aspect of rationality, they are judgements about *reasons*. In other words, the current followers of Aristotle are not only the most prevalent proponents of virtue theories, but also tend to be external reasons theorists, even though they might not all explicitly declare themselves to be such. This is why it is important to determine whether, by introducing the concept of virtue to our account, we have also introduced a source of external reasons.

We could attempt to sidestep this problem by claiming that as we have adopted a specifically Humean understanding of virtue, we need not therefore consider the implications of the Aristotelian account just because we happen to use similar terms. However, even though we do not use technical terms such as phronesis or eudaimonea, or embrace the concepts which they represent, there is enough similarity between the Humean and the Aristotelian account to raise the possibility that the Aristotelians have identified implications which are produced by any account using a recognisable concept of virtue, and that these implications lead to external reasons. Furthermore, there are aspects of the Aristotelian account which fit our everyday expectations and which accord with the universalistic intuition about reasons. So, there is an implication within the general concept of virtue that judgements of virtue are things which we can get right or wrong. For example, it is part of our common understanding of honesty that we can judge someone to be honest, yet be mistaken: we might fail to detect a lie or, more subtly, might fail to realise that an agent who acts honestly in a single instance is not revealing an honest character but is, through this exceptional act, disguising a dishonest character. Similarly, we may fail to recognise that a situation calls for honesty: the degree of honesty appropriate to those occasions on which we are asked for our opinion on a friend's new artwork, house or partner is notoriously hard to judge, but we do imagine that there is a right answer. It is the thought that virtue judgements may concern matters of fact which is contained within the concept of phronesis, which expresses the universalistic intuition and which potentially leads to external reasons. Similarly, we recognise that there are sufficiently common aspects of human life for us to have a recognisable human nature, and for our lives to go well or go badly in characteristically human ways. We even intuitively sympathise with the much contested claim that the life of someone who is materially successful yet irredeemably vicious has gone wrong in some way, even if that person appears content in his or her viciousness. Even if we do not refer to eudaimonea by name, once we begin to formalise this recognition of common forms of life and modes of flourishing by talking of virtue, we raise the possibility of external reasons.

Fortunately, we do not have to address the entire broad and rapidly growing body of virtue theory to see whether the implications of employing the concept of virtue inevitably lead us to external reasons. Two writers, John McDowell and Philippa Foot, have recently engaged directly with these questions. By considering their arguments we will find not only that these aspects of Aristotelian virtue theory do not support external reason claims, but that these theories are rather closer to our account than the external reasons theories which we considered earlier. We will start with John McDowell.

3.2.2.1 McDowell's Argument from Phronesis

The foundations of McDowell's arguments are expressed in a wide range of papers, but the way in which they relate to our discussion is fortunately succinctly expressed in a single paper which deals directly with Williams' account of internal reasons: 'Might there be external reasons?'[188] Equally fortunately, as this paper was published as part of a collection of essays on Williams' philosophy, Williams also provided us with a direct reply. Recall that the central claim of Williams' theory and of our account of internal reasons is that an agent's practical reasons are those which can be derived from his or starting motivational set by sound deliberation. Unlike most of the other challenges we have considered, McDowell is not primarily contesting the role of motivations in producing reasons. He takes Williams to be affirming the Humean claim that reason alone cannot give rise to motivations and asks, 'If the rational cogency of a piece of deliberation is in no way dependent on prior motivations, how can we comprehend it giving rise to a new motivation?'[189] McDowell's challenge is directed at the constraint imposed by sound deliberation: that is, he is asking whether an agent's reasons can be restricted to those to which he or she could deliberate. The argument goes something like this. Firstly, we accept that action cannot be explained in the absence of motivation. Secondly, we accept that reason alone cannot produce new motivations totally independently of old motivations. Thirdly, we suppose that if external reasons exist, they exist independently of the agent's motivations. And, finally, we insist that, although we accept that motivations cannot be produced by reason alone, we do not need to suppose that deliberation is the only route to the agent's external reasons. As McDowell says, 'The crucial question is this: why must the external reasons theorist envisage this transition to considering the matter aright as being effected by correct deliberation?'[190] In other words, the agent may have reasons which cannot be reached by sound deliberation, but rather through non-rational means such as conversion.

McDowell's argument that reasons exist even though they cannot be reached by sound deliberation takes the form of a challenge to Williams' understanding of reasons. His objection to Williams' account is essentially that it concentrates too closely on the explanatory dimension of reason, and does not leave sufficient room for the critical dimension of reason. In particular, he argues that the account is 'psychologistic'.[191] This term was apparently coined by Frege when arguing that the principles of logic should not be constrained by the way in which agents happen to think; that they should be 'laws of truth' rather than 'laws of thought.'[192] In this context, McDowell is using the concept to argue that the way in which agents happen to deliberate should not constrain the reasons they have. This is obviously entirely contradictory to our account of internal reasons, within which, as we have developed it, reasons have come to depend more and more on the way in which the agent deliberates. Although McDowell does not say so explicitly here, he is invoking the Aristotelian concept of phronesis, and the related concept of the

---

[188] 'Might there be external reasons?', but also see 'Are Moral Requirements Hypothetical Imperatives?' and 'Virtue and Reason' in *Mind, Value and Reality*.
[189] 'Might there be external reasons?', page 72.
[190] 'Might there be external reasons?', page 72.
[191] 'Might there be external reasons?', page 77.
[192] 'Might there be external reasons?', page 77.

phronimos: the agent who has achieved the virtue of practical wisdom as well as the other virtues. Williams recognises this when he says that, for a certain type of external reasons theorist, ''a correct deliberator' means someone who deliberates as a well-informed and well-disposed person would deliberate,' and that, 'in McDowell's account, this seems to be someone like Aristotle's phronimos, or, as McDowell puts it, someone who has been properly brought up.'[193] In other words, the idea of phronesis is important here because McDowell supposes that the reasons we have which properly represent the critical dimension of reasons are those which would be recognised by the phronimos, and any current inability to reason in the same way as the phronimos represents a deficiency in our capabilities rather than a negation of those reasons. Furthermore, McDowell does not suppose that we must be capable of reasoning ourselves into the position of the phronimos. He allows that there may be all sorts of ways of getting us to the attainment of phronesis through means which, although they could not straightforwardly be described as non-rational, nevertheless do not involve a direct exercise of reason: 'The idea of conversion would function here as the idea of an intelligible shift in motivational orientation that is exactly not effected by inducing a person to discover, by practical reasoning controlled by existing motivations, some internal reasons that he did not previously realize he had.'[194] In summary, then, McDowell is arguing that the critical dimension of rationality is capable of providing us with reasons that are not dependent on the particular motivations of an agent, and that action in accordance with those reasons can be explained by the possibility of the agent coming to acquire both the ability to recognise those reasons and the accompanying motivations through means other than deliberation.

I do not propose to follow Williams' response to McDowell's paper exactly, but rather to concentrate on three points of particular relevance to our discussion. Firstly, Williams points out that his account provides for the critical dimension of reasons by allowing that an agent's reasons may not be simple and clear, but may rather be complex and obstructed by obstacles: 'But this does not imply . . . that the agent should be able to conduct the deliberation in fact. Perhaps some unconscious obstacle, for instance, would have to be removed before he could arrive at the motivation to ϕ.'[195] The possibility of deliberative obstacles and their influence on the existence of reasons and how we talk about them is, of course, something that we have discussed at length. We have given more formal voice to Williams' thought about obstacles by drawing our distinction between strongly internal and weakly internal reasons. If our account did not afford a critical faculty to reason then we would only acknowledge the existence of strongly internal reasons; but, of course, we allow that weakly internal reasons exist as well.

The second relevant point within Williams' response is his acknowledgement of McDowell's accusation of psychologism, and his rejection of the suggestion that it presents a problem for his account: 'I accept that the account is psychologistic, in the sense that on my view a statement about A's reasons is partly a statement about A's psychology. I do not see this as an objection, as it is (I agree) an objection to say that a theory of arithmetic is psychologistic.'[196] I believe that this is exactly the right position

---

[193] 'Replies' in *World, Mind and Ethics*, page 189.
[194] 'Might there be external reasons?', page 74.
[195] 'Replies' in *World, Mind and Ethics*, page 189.
[196] 'Replies' in *World, Mind and Ethics*, page 191.

to take. Any Humean account of practical reason must accept that it is psychologistic, as it is committed to insisting that reasons are dependent on motivations; and it is this very relationship to psychology which provides such accounts with their attractiveness and plausibility. However, if we need them, there are additional ways of defusing the accusation of psychologism. We can point out, as Williams does, that we are not in the business of defining the principles of logic, but of determining the basis on which contingent agents possess reasons for action in variable circumstances. The intrusion of psychology seems entirely appropriate to the latter enterprise, even if it is anathema to the former. Alternatively, we can point out that we have not argued that all of the principles of practical reason are subject to the psychology of agents. Even Hume allowed minimal criteria of truth and falsehood for practical reasons, and within our account we insist that deliberation is sound, even if we have said that the boundaries of sound deliberation are indeterminate.

The third and perhaps the most interesting point within Williams' response, however, is one which we have not considered in our discussion so far: that the phronimos as conceived by Aristotle and implied by McDowell is an ideal type of agent and, even if we accept the theoretical existence of this ideal type, then the reasons possessed by the phronimos cannot be the same as the reasons possessed by normal agents who are rather less than ideal. As Williams argues: 'Aristotle's phronimos (to stay with that model) was, for instance, supposed to display temperance, a moderate equilibrium of the passions which did not even require the emergency virtue of self-control. But, if I know that I fall short of temperance and am unreliable with respect even to some kinds of self-control, I shall have good reason not to do some things that a temperate person could properly and safely do.'[197] The significance of this point for McDowell's argument is that it throws into doubt whether he has actually discovered the potential for external reasons which belong to the agent now, in his or her current circumstances and state of virtue, or whether he has only discovered reasons which would belong to the agent if the agent was ideal. Remember, McDowell's argument was that we could suppose external reasons to exist by imagining that their apprehension and the acquisition of any related motivation to action lie on the other side of a conversion which could be achieved through non-rational means. But it is one thing to acknowledge the possibility of conversion, and another to say that this gives the current, unconverted agent reasons to act now. At the most, it might give reason to seek conversion.

Put more informally, most of us recognise that we could be better than we are, and we might even express such intuitions in terms of virtue: we might wish that we were more prudent, more steadfast in the face of temptation, and so on. We might also recognise that we are not capable of reasoning ourselves to such improvement. So, we might adopt programmes of self-improvement, such as forcing ourselves to tell the truth in all circumstances, until honesty becomes an ingrained habit. But at most, our ambitions give us internal reasons to follow the programme now, not to act exactly as our improved selves would. Indeed, it seems impossible that we could act exactly as our improved selves would, or, even if the outward appearance of our actions was the same, to act for exactly the same reasons. When I have achieved a thoroughly honest disposition, the fact of honesty is enough of a reason for me to act. When I am still struggling to achieve such a disposition, and am constantly tempted by the ease and rewards of dishonesty, my

[197] 'Replies' in *World, Mind and Ethics*, page 190.

reasons include my wish to attain an honest disposition. An example might be that of an alcoholic following the twelve step programme used by Alcoholics Anonymous. Some of the steps in this programme are decidedly non-rational, but they are intended to build habits of thought and behaviour conducive to remaining sober. The agent the alcoholic would ideally like to be has no problem staying sober, but an important part of the programme is to recognise that this can never be achieved; the alcoholic always has the reasons of an alcoholic, even if he or she never drinks again. The point here is that, even if we can imagine an ideal which we currently fall short of, and even if we can anticipate the reasons which we will have when we have attained the ideal, the reasons we have now for taking action to attain the ideal remain rooted in the agent's ambitions: they are internal reasons. If we lack both the attainment of the ideal and the aspiration to reach it, then we cannot see why that ideal gives us reasons now. It seems that the answer to McDowell's question, 'Might there be external reasons?' is, 'Not in this fashion.'

3.2.2.2 Foot's Argument from Natural Goodness

The thought that Aristotelian theories hold up an ideal, but do not give agents the same reasons as one who has attained that ideal, also applies to arguments for human nature and eudaimonea as potential sources of external reasons. Such arguments are perhaps best exemplified by the work of Philippa Foot, particularly in her last book, *Natural Goodness*. We do not have space to consider the entire argument or its variants here, but can summarise it in four steps. Firstly, the Humean distinction between fact and value, especially as it was expressed throughout the middle of the 20th century, is decried as false.[198] Secondly, in a manner which strongly reflects Aristotle's teleological view of the natural world, it is claimed that calling an entity such as an animal 'good' is exactly what transcends the distinction between fact and value. The judgement that an animal is good is taken to carry the implication that it is good *of its kind*. So, according to Foot, 'Nobody would, I think, take it as other than a plain matter of fact that there is something wrong with the hearing of a gull that cannot distinguish the cry of its own chick, as with the sight of an owl that cannot see in the dark. Similarly, it is obvious that there are objective, factual evaluations of such things as human sight, hearing, memory and concentration, based on the life form of our own species.'[199] In other words, facts about the nature of a species and about an individual member of that species determine our evaluation of that individual. Thirdly, the nature of the human species includes all those characteristics such as sociability and mutual dependence which give rise to virtues such as benevolence, institutions such as making and keeping promises, and, if we are to believe Aristotle, the location of eudaimonea in the virtuous life. To Foot, then, the facts of human nature determine those things which we judge to be good about humans, and the facts of the conformance of an individual to that nature determines our judgement of the individual. If honesty is part of the natural end for our species then discovering someone to be dishonest means that we must judge that there is something wrong with him or her. We can already see how, if judgements of value depend on fact, reason plays a role in evaluation within Foot's scheme. However, the connection is made even stronger by the fourth step in the argument, in which it is insisted that the proper exercise

[198] For example, see chapter one, 'A Fresh Start?' in *Natural Goodness*.
[199] *Natural Goodness*, page 24.

of reason is part of the proper end for humans; to fail to reason correctly on the basis of the good for humans is not only to go wrong in reasoning, it is to offend against that good. Of course, if reasons are dependent on a universal human nature rather than on the contingent nature of individuals, then they are external reasons, and the possibility of the existence of these reasons is raised as soon as we invoke the concept of virtue.

There are two ways in which we can respond to this argument. Firstly, we can draw from one of the responses made by Williams to McDowell's argument for external reasons: that the person who has not attained the same level of virtue and practical wisdom as the phronimos does not have the same reasons as the phronimos. If the natural good of humanity is found in virtues such as honesty and benevolence in an analogous way to the natural good of animals being found in conformity to their biological, instinctively prescribed forms of life, then most of us fail to attain this good in some way. Most of us may tend to honesty much of the time, but we do not all attain a thoroughly internalised sense of honesty in quite the same way as a gull chick possesses the instinct to cry for food, or even in the way we possess our own instincts and responses, such as hunger, pain and some forms of fear. This is especially the case when we consider the more exalted virtues such as benevolence and courage, which many of us aspire to but fail to attain, and which some of us fail even to aspire to. Questions of virtue and vice would not be as important as they are unless we often fell short of virtue and stumbled into vice. So, as Williams did of McDowell, we must ask whether, even if we accept that there is a determinate human nature which contains an ideal of human goodness, this is capable of supplying common reasons to all humans, or whether those reasons may be modified or even negated by the contingent nature and circumstances of the individual agent. Of course, this response does not rule out the possible existence of external reasons based on human nature; all that is required for external reasons is that they apply independently of motivations, not that the same external reasons apply to everybody, however much that may seem to be the tendency of external reasons theories.

However, the thought that we may fail to attain an ideal of human nature, and that this has implications for any reasons derived from that ideal, prompts our second response to Foot's argument, in which we point out that the concept of human nature is a perennial source of controversy, and is surrounded by wide-ranging arguments which touch on topics as varied as genetics, politics and sexuality. This controversy is only exacerbated by the connection Foot attempts to establish between human nature and judgements of the good for humans. I tend, unsurprisingly for someone arguing for an essentially Humean theory, to believe that there is such a thing as human nature, and that it is largely common between individuals, even given phenomena such as vast and persistent cultural differences. However, this is not the same thing as supposing, as Foot and other followers of Aristotle seem to do, that there is a sufficiently uniform human nature to sustain an ideal of fulfilment which is capable of providing us with determinate reasons independent of our motivations. The difficulties associated with such a concept are apparent enough when we consider subtle variations between individuals, but become even more glaring when we think of those people who are unfortunately impaired in some way. The potentially thoroughly fulfilled life of someone who happens to be disabled makes us balk at the idea that the ideal of human flourishing can only be attained by the able-bodied, and also at the idea that such a person somehow fails to attract the judgement 'good'. It could, of course, be argued that such a person would be better off if

he or she was not disabled, but being better off is not the same as flourishing or attaining fulfilment. Furthermore, talk of whether someone would be better off or not without a disability misses the fundamental lesson of such an example: that our experience and our expectations tell us that there are many ways to flourish and to be 'good'.[200]

I do not think that we have done enough here to show that Foot's, McDowell's, or any other Aristotelian argument must necessarily fail; but we have done enough to show that the path from virtue to external reasons is not straightforward and is certainly not inevitable. However, we have also seen enough to realise that Aristotelian theories do not seem to be in conflict with all aspects of our own theory; certainly not to the extent of the other external reasons theories we have considered. Furthermore, some aspects of Aristotelian theories fit our experience: we do aspire to ideals and we do give ourselves reasons that accord with those aspirations. These thoughts indicate how best to engage with Aristotelian theories. As Rosalind Hursthouse admits unlike the theorists we considered earlier, neither Aristotle nor modern Aristotelians pretend to be arguing from a position wholly outside ethics: "The pretensions of an Aristotelian naturalism are not, in any ordinary understanding of the terms, either 'scientific' or 'foundational'. It does not seek to establish its conclusions from 'a neutral point of view'. Hence it does not expect what it says to convince anyone whose ethical outlook or perspective is largely different from the ethical outlook from within which the naturalistic conclusions are argued for.'[201] I do not want to underplay important theoretical differences such as the concept of eudaimonea as the proper end for humans, but it seems to me that, if we could set such differences to one side, the best way for us to deal with Aristotelian virtue theorists within our account is as extremely optimistic internal reasons theorists. So, if we return to the example of McDowell, if what he means is that the example of the phronimos gives us reasons to exercise our capacities to emulate this ideal because we have common dispositions best served by achieving that state, then we have no problem; as long as we can also allow that is at least conceivable that someone could lack the dispositions and therefore lack the reasons. Of course, McDowell is saying rather more than that, and whether this is really the right way to deal with his argument is a question to be addressed in rather more depth elsewhere. For the time being, though, I believe that we can say that the onus is on the Aristotelians to show that virtue leads to external reasons, rather than on us to show that it does not, and that consequently we can continue to employ the concept of virtue without supposing that by doing so we are compromising our account.

---

[200] We may also note that recent and foreseeable advances in medical technology make this concept even more troublesome. If someone is genetically or surgically enhanced to be faster or stronger or to have more acute senses than the normal run of humanity, we must ask whether, under Foot's scheme, we should regard that person as defective because of their variation from normal human nature, as belonging to a kind of his or her own, or as raising the standard of human natural goodness for the rest of us.

[201] *On Virtue Ethics*, page 193.

3.3 Reasonableness

So, we can now consider whether we have found the components of a Humean virtue. We shall start by considering whether the motivations, attitudes and assumptions which we found expressed in everyday talk of reasons and the behaviour they produce generally enjoy our approval. To recap, we found an aversion to arbitrariness, an aversion to motivations which disrupt the apprehension and sincere expression of reasons, a desire for intelligibility, and a hope that we can reach conclusions capable of producing agreement, and we grouped them together under the heading of a desire for sustainable reaons. Fortunately, it is not hard to discover what our attitude is towards such phenomena, as our positive regard for them is immediately apparent from our ordinary behaviour. When we are confronted with the question of whether we would prefer an agent to seek sustainable reasons and to act in accordance with them, or to disdain such reasons by failing to look for them, or by, having found them, choosing to act against them, our instinctive preference is for the former.

However, we can do more than simply insist on the obviousness of our positive regard for the common motivations concerning reasons. As well as giving us reasons which meet our expectations in particular instances of deliberation, possession of these motivations and behaviour in accordance with them make us into reliable partners in deliberation and action. The reasons of someone who is averse to arbitrariness are more likely to be based on considerations which we and they consider to be relevant to the matter at hand. If someone is suspicious of the disruption to deliberation that can be caused by strong motivations, or of the temptation to present one's reasons insincerely, then we can be more confident that what that person professes his or her reasons to be really are his or her reasons. An agent who wants reasons to be intelligible is more likely to offer reasons that we can understand and, if we are drawn into a debate about those reasons, is more likely to be able to present us with deliberation that we can understand. And someone who wants reasons to be sustainable is more likely to act for reasons which have an appropriate degree of justification, and is also likely to respond to challenges to reasons, either by seeking more robust justification or by accepting the need to seek new reasons. In short, the people who possess and heed the motivations, attitudes and assumptions we have considered are those whose reasons we can understand and trust. We can engage with them in practical matters with the expectation that they will either share our reasons, or that if they do not share our reasons, we will have some common grounds to resolve our disagreement. We must also recognise that this sort of practical engagement does not only happen with others; it also happens within ourselves. Our motivations and deliberation are not always so explicit that we are fully aware of them, and we sometimes find that we have to pull ourselves up and work out why we are doing what we are doing and why we think that we have particular reasons for action. When we do this we hope that we will find that our reasons are sustainable.

Of course, it takes more than just a group of motivations and behaviour that we approve of to constitute a virtue. To find such a virtue it must be possible to identify these motivations and the behaviour they produce as constituents of a character trait, and preferably one that we can recognise from our everyday lives. I believe that we can identify such a familiar character trait: *reasonableness*.. Although reasonableness is not cited in the many lists of virtues compiled by writers working in this area, once we have

suggested that reasonableness is a virtue it seems intuitively plausible: reasonableness is a character trait that we recognise, and it is one that we value. In our everyday talk we often apply this term to people ('I'm a reasonable man.') but its character is most clearly revealed by the way in which we seem to use it most frequently: as an indication of agreement to a proposal for action ('That seems reasonable to me.').

To understand what we usually mean by reasonableness when we use the term casually with respect to agreements we can imagine a man attempting to come to an agreement about action with a friend, possibly about something as mundane as which film to go and see that evening, and ask what it would take for the agreement they reach to be judged reasonable in ordinary terms. Whatever agreement they reach would certainly have to be good enough to satisfy both of them. They could of course reach an agreement which satisfied only one of them, because the other was browbeaten into submission, or simply decided that the decision was not worth an acrimonious dispute. But such an agreement would not be judged reasonable; it would be better described as a somewhat arbitrary compromise. Reaching a reasonable agreement would also naturally involve a degree of understanding, both of the other's motivations and of his or her deliberation. If the man does not want to see *The Maltese Falcon* because he thinks it is a nature film and he wants to see a film noir, then the agreement will not be reasonable if this misunderstanding is not made explicit. Similarly, the agreement will not be reasonable if it is not what it appears on the surface; that is, if one of those people involved makes an agreement because he or she knows that it will not work out in the way overtly intended, as it has been insincerely framed in the service of some hidden motivation. If the man agrees to go to the cinema to see *The Maltese Falcon* in the knowledge that tonight the cinema is showing a different film then the agreement can only seem reasonable until it is discovered that it is not so. However, this only means that reasonable agreements are incompatible with insincerely expressed or hidden motivations, not that they are incompatible with all motivations; indeed, without motivations there would be nothing to agree or disagree about, and the most reasonable agreements will be the ones in which these motivations are most sincerely expressed. Finally, a reasonable agreement is one which has been reached through some sort of shared deliberative process. If the man simply insists on going to see the film he wants to see without allowing any room for discussion, there is no possibility of a reasonable agreement. So, reasonable agreements are those which are not arbitrary, are not distorted or insincere, but nevertheless reflect the motivations of those involved, and are intelligible; in other, words, those which are based on sustainable reasons.

We can extend this understanding to the character of agents. Because of the association of reasonableness with agreements we may primarily think of someone who is judged to be reasonable as someone who is prepared to accept compromises. However, this does not mean that we think that someone who is reasonable is someone who always gives way when challenged, or someone who is always prepared to sacrifice his interests for the sake of others. Rather, even in our intuitive association of reasonableness with compromise, we recognise that someone who is reasonable has a position to compromise *from*, just as a reasonable agreement is one which offers some satisfaction to the various parties involved. Reasonableness is not mere capitulation. We expect that someone who is reasonable is someone who in practical matters is capable of seeing that sometimes it is just as important to reach a conclusion about action as it is to defend a particular position.

This does not mean that compromise, agreement or conclusion can always be reached by the reasonable agent, as some things are not open to challenge, but it does mean that the reasonable agent is capable of seeing when the satisfaction of some motivations is better than the satisfaction of all motivations, because only the former is on offer. As ever with talk of motivations we must remember that we are not just talking about the efficient satisfaction of self-interest; the reasonable agent may be a peace-keeper negotiating, at great personal risk, the safe passage of refugees at the expense of surrendering control of a town. So, our informal expectations of a reasonable agent are that such an agent is someone who is committed to certain practical positions, but who is also committed to reaching conclusions about action, and is further committed to reaching such conclusions in a manner which is governed by reason. That last commitment is particularly important. After all, a literal interpretation of 'reasonable' might be 'capable of being governed by reason.'

These commitments, to specific practical positions, to the achievement of conclusions, and to reasoned deliberation, mean that we also have further expectations of the character of reasonable agents. Someone who is consistently characterised as reasonable may also be expected to be reliable and trustworthy (and possibly even somewhat dull; the term reasonableness carries no hint of passion). Perhaps most intriguingly, we may also think of someone who is genuinely reasonable that he or she is capable of transmitting that reasonableness to others, even if only temporarily; I am sure that we can all think of someone who, through sheer persistent application of reasonable argument, calms down the most passionate debate. We may even find such an effect irritating from time to time, but that is most likely when part of the purpose of our passionate debate was to have a passionate debate, and the reasonable agent who patiently leads us to a practical conclusion has rather missed the point. Once again, we can also express these informal thoughts about the character of the reasonable agent in terms of the motivations we have found concerning reasons and deliberation. So, we can say, on our ordinary understanding, that a reasonable agent is one who is uneasy about admitting arbitrary considerations into practical deliberation, who sincerely expresses his or her motivations concerning the matter at hand, who can avoid or regulate the influence of strong motivations on the processes of deliberation, even if those motivations are the driving force behind that deliberation, whose deliberation we can follow, even if this requires some explanation, who has a position but is prepared to move from it, either to achieve a conclusion or if the position is sufficiently undermined, and who has a respect for deliberation, to the extent that this respect may be contagious.

So, I think that we can regard reasonableness as a virtue, and can define it in terms of the general heading under which we grouped our common motivations, attitudes and assumptions concerning reasons; we can say that:

> *reasonableness is the embedded tendency of an agent to pursue sustainable reasons, to be reliably successful in finding those reasons, and to act on those reasons when they can be found.*

By now we are familiar with most of the terms in this definition, but it is worth saying a little more about the terms 'pursues' and 'reliably successful'. I have used the term 'pursues' for four reasons. The first is quite straightforward; the term acknowledges

that an agent may be reasonable, but nevertheless may fail to find sustainable reasons within a particular practical context. Sometimes this may be because the agent is incapable of finding the sustainable reasons which do exist, and sometimes it may be because such reasons just don't exist. The important point, though, is that the reasonable agent tries to find sustainable reasons, even if this attempt is fruitless, and does not fool him or herself into thinking that such reasons have been found when they have not. The second reason for using the term 'pursues' is that it does not constrain what it involves in the attempt to find sustainable reasons; it is an open-ended term. As we have seen, there are many different ways of going about trying to find reasons, and part of being reasonable is using all of those means which seem appropriate to the practical situation. So, the pursuit of sustainable reasons may involve profound, introspective, solitary thought, or it may involve more gregarious activities such as a group of friends playfully exchanging ideas about what to do. The third reason for talking of pursuit is that, just as physical pursuit of a quarry is an activity that can be frustrated and misled; the pursuit of sustainable reasons may involve acknowledging that sometimes reasons we thought were secure have been undermined by challenges which are irksome but nevertheless valid; sometimes a pursuit which we thought was over turns out not to be finished. The reasonable agent may find that cherished conclusions, reached after a difficult deliberative struggle, are thrown into doubt by a chance thought or observation. This does not mean that all challenges to reasons will be accepted, but that the reasonable agent is the agent who recognises when a reason that seemed sustainable is shown not to be, and also realises that it is time to start deliberating again, however tiresome that may be. The final reason for using the term 'pursues' is simply that it emphasises the motivational roots of the virtue; those things we pursue are those things that we care about catching, and the reasonable agent cares about finding reasons which are sustainable.

In one sense, the term 'reliably successful' has similar implications to the term 'pursues'; that is, by standing in contrast to alternatives such as 'always successful', it indicates that we do not expect that agents will always be able to find sustainable reasons. However, the inclusion of the term within the definition also indicates that we expect the reasonable agent to be capable of finding sustainable reasons most of the time; an agent who consistently set out to find such reasons but failed to do so might be considered to aspire to the virtue of reasonableness but not to have attained it. However, we must be cautious about the emphasis we place on success with respect to virtues. Zagzebski argues that reliable success is a component of all virtues, saying, for example, that, 'A kind, compassionate, generous, courageous or just person aims at making the world a certain way, and reliable success in making it that way is a condition for having the virtue in question.'[202] While this claim is basically correct, we must also acknowledge that the degree to which the exercise of a virtue must be successful before the agent can be considered to possess that virtue depends on context. For example, we may say that the success of benevolence is manifested in the material difference an agent makes to the lives of others, and then realise that sometimes people we would not hesitate to describe as benevolent find themselves in situations where they are frustrated in making this difference: consider a doctor, helpless in the face of a ravaging plague, who can do nothing to ease the suffering that surrounds him. So, all we expect of virtue is reliable success, rather than perfect success.

---

[202] *Virtues of the Mind*, page 136.

Although we have offered a conveniently brief definition of reasonableness, we must also recognise that virtues resist precise definition, and that our understanding of them is best acquired through acquaintance, experience and attainment. We can get a better grasp of reasonableness, though, as well as reinforcing our claim that it is a virtue, by considering our experience of its corresponding vice: *unreasonableness*. Just as with reasonableness, although unreasonableness is not usually included in catalogues of vices, once it has been proposed as a vice it seems intuitively plausible that we should accept it as such. Indeed, the feeling of recognition we get at the thought that unreasonableness is a vice is even greater than that we get at the thought that reasonableness is a virtue. As I shall argue later, we do not typically include reasonableness in our lists of virtues because it is often a relatively modest accomplishment; we find it easier to be reasonable than we find it to be courageous, or benevolent, or sometimes even honest. By contrast, and possibly because of the common attainment of reasonableness, genuine cases of unreasonableness stand out. If we can think of examples of people who we would call unreasonable - and, unfortunately, I think that most of us have to deal with such people from time to time - we have little doubt about what it is that makes us call them unreasonable, and we have even less doubt about how we feel about them; unreasonableness doesn't just engender an attitude of disapproval, but often produces teeth-gritting, fist-clenching frustration and anger.

However, despite this feeling of recognition and common reaction to examples of unreasonableness, this candidate for a vice is not exhibited in just one way. Rather, we can identify at least four distinct manifestations of unreasonableness. The first of these is simple *laziness* in the pursuit of reasons. As we have seen, the pursuit of sustainable reasons does not demand that exhaustive deliberation is undertaken in response to every practical question, but it does require that for certain tough practical questions tough deliberation is undertaken. Sustainability does not mean complacency; rather it often arises from the very trouble taken to reach a conclusion. The lazily unreasonable agent is one who is unwilling to undertake difficult deliberation or to contemplate hard decisions, but who instead rests content with conclusions which, even if they have some merit, are not adequate to the task at hand. A lazily unreasonable agent may be a man who declares that he is not going to vote because politicians are all the same, and are only in it for themselves. The important thing to note about this sort of deliberative behaviour is that it is not simply a case of an agent finding sustainability where we find doubt. Such differences do occur, and our debates with others about their reasons are often attempts to discover how they can be confident about a conclusion which fails to convince us. If the agent is truly exhibiting the lazy form of unreasonableness then there is an element of bad faith involved. The agent is not really confident in the conclusion he or she claims to hold; confidence is merely claimed as a way to avoid further deliberative work. The man who refuses to vote is not really abstaining out of principle or a genuine disillusionment with politicians; he just can't be bothered with the deliberative effort required to engage in the political process at all.

The second manifestation of unreasonableness also involves bad faith, but to an even greater degree that that exhibited within the lazy form of unreasonableness. That manifestation can best be characterised as *slyness*. Agents can be described as slyly unreasonable when they exploit the common conventions of deliberation and common motivations concerning reasons in order to get others to act in accordance with their

wishes, not through persuasion, but through manipulation. In short, they are unreasonable because they subvert the reasonableness of others. Slyness works because the motivations concerning reasons are common and because these motivations are engaged by the judgements of others as well as by our own judgements. So, if someone tells us that within a piece of publicly articulated deliberation we have omitted or distorted a vital piece of deliberation, we will be discomfited; we will worry that our reasons are subject to arbitrary influences, that they are no longer intelligible, that they have become unsustainable and so on. We may also become concerned that others are judging us to be acting on unsustainable reasons. Our common motivations and our attention to the judgements of others mean that challenges to our reasons and deliberation are capable of disrupting our confidence, and this can be taken advantage of by the slyly unreasonable agent.

We have already met a straightforward example of slyness: the phenomenon of bluff described by Williams in 'Internal and External Reasons', in which one person simply insists that another person has a reason, even though he or she is not aware of it, and has no motivations from which that reason could be derived through sound deliberation. It takes little more than the assertion that a reason exists to make us wonder whether it is truly so. This may be partly due to our underlying awareness of all the different ways that deliberation goes wrong; we are often uncomfortably conscious that we are fallible, and that we are subject to all of the deliberative obstacles which we discussed in the early part of this thesis. Sometimes an insistence that we have a reason which we did not previously apprehend is simply right, and on reflection we adjust our understanding of our reasons accordingly. Consequently, genuine cases of bluff may be rare; an apparent bluffer may sincerely believe that he or she is helping the agent push past deliberative obstacles. But our occasional need for this help creates the opportunity for genuine cases of bluff.

Perhaps the most insidious form of sly unreasonableness occurs when the sly agent exploits the reasonableness of others to distort their deliberation. Imagine that the committee which runs an amateur dramatics society is trying to decide what their next production might be: *Oklahoma!* or *HMS Pinafore.* As often happens in such organisations, passions are deeply engaged: two factions emerge, each of which is fervently in favour of one of the options and implacably opposed to the other. Some members of each faction attempt to be reasonable about the disagreement; they examine their deliberation to see whether they are influenced by arbitrary considerations, they sincerely attempt to make their deliberation explicit and intelligible, and they wonder whether, in the face of such committed opposition, their reasons can really be considered sustainable. Other members of the committee are not so virtuous, however. They deliberately introduce considerations into the argument which, although they resemble concerns which ought to be taken into account within deliberation are not really factors which matter within the context of this particular practical problem at all; they are simply attempts to disrupt the process of deliberation, made in the knowledge that the mostly reasonable agents who make up the rest of the committee will feel obliged to pay them at least some attention. So, an advocate of *Oklahoma!* Might suggest that it would be dangerous to stage *HMS Pinafore* at a time when the country is in the middle of an unpopular and intractable conflict, and the public might look askance at anything with militaristic connections. Similarly, an advocate of *HMS Pinafore* might argue that staging

a musical such as *Oklahoma!* in which one of the characters dies in a knife fight might exacerbate the recent tendency of local youths to carry knives. Of course, it is not inconceivable that people could be extremely sensitive to exactly these issues, and that the concerns they voice about them are entirely sincere. However, within our example we may suppose that the people who raise these issues are not sincerely concerned about them; they just offer a means of subverting the collective deliberation of the committee in order to get what they want. Public, collective deliberation is especially vulnerable to such subversion, especially when it is conducted in such an outwardly polite context as the committee of an amateur dramatics society. Because of our pursuit of sustainable reasons it is difficult to ignore challenges that are made towards our reasons and our deliberation. So, once the respective unreasonable proponents of *Oklahoma!* and *HMS Pinafore* have raised the issues of youth violence and anti-militarism it is very difficult to make those issues go away. One can almost hear the committee members groan as they realise that their deliberations have just needlessly grown more complicated, and that their previous confidence and simplicity of approach cannot be recovered without being rude.

Rudeness is almost expected behaviour from those people who exhibit our third form of unreasonableness: *autocracy.* Autocratic unreasonableness occurs when the unreasonable agent adopts a particular attitude to sustainability; he or she may still be motivated to seek reasons which are sustainable, but has come to believe that his or her approval is all that is required to achieve that sustainability. Similarly, the agent feels no constraints to make reasons intelligible to anyone except him or herself, and may even be inclined to act on reasons which he or she does not really understand; reasons which might be better described as whims. In other words, the unreasonable autocrat issues edicts rather than offering proposals for action; such agents have an unwarranted excess of confidence in their reasons, to the extent that they are prepared to brook no challenge. This phenomenon may look like the variety of solipsism Nagel saw as the consequence of failing to recognise others as sources of reasons. There is a definite resemblance between autocracy and solipsism as understood by Nagel, but there is also an important difference; Nagel's solipsism seemed to arise from within the agent, as a mistaken understanding of the nature of agents and their reasons, whereas autocratic unreasonableness often arises through the immersion of the agent in institutions of power which simultaneously grant the agent authority while preventing that authority from being challenged.

Such institutionalised unreasonableness reaches its obvious peak in despotic regimes, and figures such as Caligula are exemplars of unreasonableness; whims of the moment are taken as foundations of dramatic action, and challenges to those whims are taken as treason.[203] It is something of a cliché for writers to place the words, 'I'm a

[203] Of course, we have encountered frivolous whims before, and defended them against accusations of irrationality; Hume famously said that, "Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger' (*A Treatise of Human Nature*, Book II, Part III, Section III, 463). The important thing to remember here is that Hume is arguing that the *preference* would not be *irrational*, not that action on that preference would not be unreasonable in our sense. In most people this particular preference would not exist, and even if it did, it is a long way from the preference to conclusions about action, through sound deliberation and the consideration of many other preferences. The point about unreasonable autocrats is that they do not accept the same constraints on their reasons as other people do, and, if they belong to the variety of autocrat we might meet in James Bond stories, may progress from trivial desire to global destruction on the basis of thoroughly bad reasons, even with

reasonable man,' in the mouths of such despots, usually at a point when they have a struggling victim within their power. These words are intended to be chilling because what comes next is extremely unlikely to be reasonable; the despot's power makes it unnecessary for him to be reasonable, and he is far more likely to propose a devastating compromise to his victim's integrity in exchange for a satisfaction of some desire than to seek to become a reliable partner in action. However, this form of unreasonableness does not only manifest itself in such extreme forms as that exhibited by the rulers of despotic regimes. It is also found in everyday situations, particularly in the workplace; it seems that we are particularly vulnerable to the temptations of power, and those temptations include the opportunity to avoid having to justify our reasons or make our deliberation intelligible.

We should note, though, that autocratic unreasonableness does not simply constitute the exercise of authority, or inducing people to act without a full explanation. There are many occasions on which group efforts must be co-ordinated, and can only succeed if members of the group cede authority to a leader who takes responsibility for decision making, but is not practically able to explain all of those decisions at the same time as realising the goals of the group. We may think, for example, of the captain of a sports team who has to make quick decisions about tactics and positioning, but cannot operate if these decisions must always be explained to the team members during a game. This authority gives rise to autocratic unreasonableness when the original reasons for not offering explanation and justification are forgotten, and the exemption from the conventions of reasonable discourse become associated with the role, or even worse, with the person. Such autocratic unreasonableness can often be particularly frustrating because, as mentioned earlier, the conditions which give autocrats power often also protect them from criticism and challenge. Someone who is subject to the autocratic form of the vice of unreasonableness is not only likely to reject the offer of help to clear away deliberative obstacles and apprehend reasons clearly; they may even have people whose job is to prevent them from even hearing such offers.

The final form of unreasonableness we will consider differs from the others in that it is produced by an excess of the motivations concerning reasons and deliberation rather than a deficiency. We can use the label *pedantry* to indicate the placing of an excessive burden on sustainability, to the point where it is more closely resembles the certainty sought by external reasons theorists, and where no reasons which we would normally accept seems good enough. Pedantry in the service of reasons occurs when the agent cannot achieve confidence that any reason suits a particular situation, and picks away at the basis for reasons in search of something more fundamental, with the effect of undermining his or her own confidence, as well as that of others. The obvious candidates for people who suffer from this form on unreasonableness are our external reasons theorists; or, indeed, anyone who worries at reasons to the point of plunging into the philosophical gloom in which Hume sometimes found himself.[204] However, I believe that the question of whether such people are really unreasonable is rather more complicated, and I shall consider it at greater length later. A more familiar example of this sort of unreasonableness is the incessant curiosity exhibited by some children; not just on questions such as, 'Why is the sky blue?' but also in the repeated, 'Why? Why? Why?'

---

respect to their own motivations.

[204] *Treatise*, Book I, Part IV, Section VII, page 316.

which follows any explanation we offer. It is surprising and sometimes alarming how quickly we run out of such explanations, or at least explanations that we can offer convincingly, and how much of our interaction with the world is dependent on trust and faith. Of course, it is unfair to judge that childish curiosity is genuinely unreasonable. Indeed, it may be seen as charming, and it is often regretted that we lose a childish sense of wonder as we mature. However, we must recognise that such views have a sentimental aspect, and that we would regard an adult who displayed such incessant curiosity while attempting to attend to practical matters as irritating, frustrating and, of course, unreasonable. As we grow in responsibility for our own actions and the actions of others we need to find confidence somewhere and continuous questioning can come to seem like self-indulgence. There are places in our lives for radical curiosity and the questioning of assumptions that we would ordinarily think of as unquestionable; and one of these places is the pursuit of philosophy. However, we must recognise that when we are engaged in endeavours which require us to radically question our most basic assumptions we are not engaged in the same sort of endeavour as when we try to make ordinary, or even extraordinary, decisions about action. Part of being reasonable is knowing which modes of enquiry to apply to which situation.

The existence of these different of forms of unreasonableness allows us to supplement our earlier definition of reasonableness as the pursuit of sustainable reasons by further defining it in relation to the vices it avoids. The requirement that reasons are sustainable means that the reasonable agent resists pressure from two directions: he or she resists our tendency towards laziness by refusing to accept reasons which are inadequate; and he or she resists our tendency towards pedantry by allowing that we can and should accept reasons once they are sustainable, even if there are other lines of enquiry which could conceivably be pursued. The thought that we have a virtue which steers us between the extremes of laziness and pedantry with regard to action has been echoed with regard to belief by Linda Zagzebski:

> One must be neither too sanguine in one's convictions nor too obsessed with the desire to inquire further before reaching the state of settled belief. One must, in short, know when to stop, but also when to start and to continue. The virtue that is the mean between the questioning mania and the unjustified conviction has no simple name, as far as I know, but it is something like being both properly inquiring and properly doubtful.'[205]

Of course, I think that, within the field of practical reason we can give that virtue a name, and that name is reasonableness. And if we want to give it a description similar to Zagzebski's we can do so: it is the mean between laziness and pedantry in questions of practical reason; something like being both properly demanding of reasons and properly accepting of reasons which are sufficiently justified.[206]

---

[205] *Virtues of the Mind*, page 154.

[206] Although he is not writing about virtues, a similar thought is captured by John Dewey in *How We Think*, when he says that, 'To take too much pains in one case is as foolish – as illogical – as to take too little in another. At one extreme, almost any conclusion that insures prompt and unified action may be better than any long delayed conclusion; while at the other, decision may have to be postponed for a long period – perhaps for a lifetime.' *How We Think*, page 78.

Despite the existence of the vice of unreasonableness, and the achievement constituted by its avoidance, the concept of reasonableness may still seem rather too mild to qualify as a virtue. As an everyday term reasonableness has tones of constraint and compromise, and when compared to other virtues such as benevolence and courage, which may be exhibited in conspicuously noble and dramatic ways, it may appear a little timid. However, there are other virtues which we do not always think of as particularly heroic. Honesty, for example, is something which we expect to exhibit and see exhibited in most agents most of the time, just as we expect that everybody should be able to attain some degree of reasonableness. However, even honesty has its moments of drama and heroism. Consider the fable of the Emperor's new clothes, in which the boy who points out that the Emperor is naked is doing no more than being honest, yet is doing something that no-one else would dare. Honesty may also rise above the mundane when it involves the recognition and proclamation of dishonesty. We are often too embarrassed to declare that someone is a liar, even when we know and he or she knows that this is the case, and even when keeping quiet makes us complicit in the dishonesty. Overcoming this embarrassment and speaking out is partly a matter of courage, but it is also an expression of the love of truth that motivates honesty. Reasonableness can produce similar moments of drama and, as with honesty, these typically come when the agent defies consensus or refuses to tolerate the corresponding vice.

What counts as a sustainable reason may vary from person to person and culture to culture, and is at least partly a consequence of deliberative habits ingrained in the culture and the individual. However, as we have seen, sustainability does not mean complacency; the term was deliberately chosen because it carries a connotation of stability rather than permanence, and our account allows that, from time to time, appropriately critical thought will undermine what was previously seen as sustainable. This means that reasonableness, mild as it may seem, may be a component of radical social change. I think that it is plausible to claim that the slow collapse of prejudice in some societies is constituted in part by a realisation that ways of thinking and behaving in relation to certain social groups are no longer sustainable. This does not necessarily mean that the motivations underpinning reasonableness initiate the change in attitudes; that may be brought about by means as various as campaigns of protest, changes in language and simply living alongside one another. The important thought is that, whatever begins the shift in attitudes, there comes a point when the reasonable agent realises that certain reasons and certain habits of thought can no longer be sustained; and the reasonable acknowledgement of this change may be both hard and praiseworthy, especially in comparison to the unreasonable alternative of denial. On a more individual level, the identification of unreasonableness may be as courageous and uncomfortable as calling a liar a liar. This is especially the case when unreasonableness takes the autocratic form we discussed earlier, and the unreasonable agent is in a position of authority over the reasonable agent. In calling unreasonableness what it is, the reasonable agent is not only undermining that authority, but the very way in which the autocratic agent thinks and behaves; imagine a junior minister in the British government telling the Prime Minister that not only are the policies he or she is being asked to implement wrong, but the reasons for implementing them are unsustainable and worse, that the thinking behind them is unintelligible. I do not wish to overstate the case; most of the time most us can be expected to be more or less reasonable in an undramatic way, just as most of us can be

expected to be more or less honest. But there are occasions when reasonableness is heroic, and this supports the idea that it is worthy for consideration as a virtue.

The possibility that reasonableness can sometimes be heroic, and the existence of the vice of unreasonableness, both indicate that what we have found in reasonableness is a genuine virtue, rather than a mere skill associated with the activity of deliberation. In *Virtues of the Mind*, Zagzebski discusses the distinction between virtues and skills at some length, identifying several different ways of making this distinction, including the need for virtues to be actively expressed in action, the intrinsic value of virtues compared to the instrumental value of skills, the possibility that virtues can be faked, and the existence of vices as the opposite of virtues. [207] I believe that reasonableness qualifies as a virtue on all these tests: as defined here, reasonableness involves not only the reliably successful pursuit of sustainable reasons, but action on those reasons; we characteristically approve of the display of reasonableness for its sake, as well as any desirable outcomes which reasonable behaviour may produce; the existence of the sly form of unreasonableness shows that reasonableness can be faked; and the existence of unreasonableness more generally shows that reasonableness has a corresponding vice.

However, the distinction between skills and virtues does not mean that deliberative skills are irrelevant to the virtue; on the contrary, virtues and skills are closely related, and particular virtues are associated with particular skills. This is certainly true in the case of reasonableness. In the early part of this thesis we talked about deliberative capabilities, and how the lack of these capabilities constituted deliberative obstacles. We can now recognise that some of those capabilities are skills associated with the virtue of reasonableness. We did not attempt to list all deliberative capabilities earlier, and I do not propose to present an exhaustive taxonomy of the skills associated with reasonableness here; this is partly because it can be difficult to know when one skill should be regarded as distinct from another, partly because some skills may also be associated with other virtues, but mainly because I believe that any such attempted taxonomy would inevitably be incomplete and inaccurate, and would also be of dubious value. I believe that it is enough that we can readily recognise that skills associated with individual deliberation, such as the ability to determine which people and which information to trust, the ability to exercise the imagination to an appropriate degree and the ability to maintain clarity about the deliberation which has been undertaken, are all closely related to reasonableness, as are skills associated with collective deliberation, such as the ability to articulate one's position clearly, the ability to understand the arguments of others, and the ability to participate in fruitful discussion. All of these skills help us to find sustainable reasons, and to make the case for these reasons to others. And the relationship between these skills and the possession of the virtue of reasonableness is that suggested by the nature of the distinctions drawn between them; we can be reasonable without possessing these skills, but without them that reasonableness may be ineffectual.

So, we have established that reasonableness is a virtue, with a corresponding vice and a supporting set of skills. And, as we have defined it as a virtue in Humean terms, we have one more job of classification to do: we must determine whether it is a natural or an artificial virtue.

---

[207] See *Virtues of the Mind*, 2.4, 'Virtues distinguished from skills'.

3.3.1 Reasonableness: A Natural Virtue?

As we have seen, Hume distinguishes between those virtues which we can expect to be valued by all cultures at all times, and those which, 'produce pleasure and approbation by means of an artifice or contrivance, which arises from the circumstances and necessity of mankind.'[208] This distinction matters because, despite what we can discover through our theorising, if we share Hume's understanding of virtue we will look to our experience and judgements to tell us about virtue in the first instance. If we know that a virtue is artificial then we know that there are circumstances in which the virtue may not be manifested, not just within individuals who lack that virtue or who are subject to the corresponding vice, but within an entire culture. If we know that a virtue is natural then we may expect to find it exhibited in some way in every culture and circumstance, and also know that if we do not find it easily then it is either manifested in a form with which we are unfamiliar, or that we are incorrect in regarding it as a natural virtue; either way, we may deepen our understanding of the virtue. I believe reasonableness is a natural virtue, but one whose manifestation varies from place to place and culture to culture, because reasonableness remains the same while what is considered reasonable at particular times and places changes.

To show this we will use a method which was employed in a basic form by Hume, but which has been used by many other writers including, recently, Bernard Williams: the method of *imaginary genealogy*.[209] As we have seen, in order to establish that justice was an artificial virtue Hume imagined a fictitious state of nature in which the benefits and disadvantages of his civilised society were absent but the basic characteristics of human nature persisted.[210] The assumption is that, if we can see how the patterns of motivations and behaviour constituting a virtue develop and become reinforced through judgements of approval in that state of nature, even if we vary what we consider that state of nature to contain, then we have discovered a natural virtue. If we can see that the virtue would only have developed under particular, specialised circumstances, then we have discovered an artificial virtue. Of course, those particular, specialised circumstances may be an inescapable part of the world we live in, meaning that the virtue, though artificial, will appear in all actual human societies; thus Hume is able to classify justice (as he understands it) as an artificial virtue, at the same time as maintaining that it will be universally encountered and valued. In *Truth and Truthfulness*, Williams uses a similar approach[211] but gives it rather more formality, and instead of describing the state of nature according to what 'poets have invented',[212] says that, 'In the State of Nature there is a small society of human beings, sharing a common language, with no elaborate technology and no form of writing.'[213] However, despite this less fanciful definition, Williams is clear that in considering the State of Nature we are never considering a real historical situation: 'The State of Nature story is a fiction, an imaginary genealogy, which proceeds by way of abstract argument from some very general and, I take it, indisputable

---

[208] *Treatise*, Book III, Part II, Section I, page 529.
[209] For Bernard Williams' employment of this method see *Truth and Truthfulness*, especially chapter 2.
[210] See *Treatise*, Book III, Part II, Section II.
[211] Williams also acknowledges the work of other recent writers who have used a similar approach, notably Edward Craig in his book *Knowledge and the State of Nature: An Essay in Conceptual Synthesis.*
[212] See *Treatise*, Book III, Part II, Section II, page 545.
[213] *Truth and Truthfulness*, page 41.

assumptions about human powers and limitations.'[214] Despite this artificiality, Williams is able to use the State of Nature to demonstrate how his two virtues of truthfulness, Accuracy and Sincerity, inevitably develop out of the material demands of the situation and the existence of differences in knowledge as basic as what Williams calls 'positional advantage',[215] which might be as straightforward as being up a tree while others are on the ground. This means that, in constructing the State of Nature we need make no attempt to ensure historical accuracy, but we must make sure that we only imbue the inhabitants of our story with the powers and limitations which humans indisputably possess. We should note, however, that a requirement of any imagined State of Nature is that it constitutes a *society*; even if we suppose that humans in the State of Nature would not form a stable community, we do not imagine that they comprise a group of people who have been suddenly thrown together in a primitive situation, as in the *Lord of the Flies*.

It has been suggested that, even though Hume uses an abbreviated form of imaginary genealogy within his argument that justice is an artificial virtue, his work is not compatible with this sort of approach.[216] Two charges are usually levelled at Hume in this respect. Firstly, it is claimed that his view of human nature is so uniform that it does not allow room for the development and variation which is manifest in history and which is required to make sense of genealogical accounts. This view of human nature as uniform is implied by his construction of a catalogue of virtues without considering whether a different person at a different time would construct a different catalogue, and is explicitly expressed when he makes claims such as, 'It is universally acknowledged that there is a great uniformity among the actions of men, in all nations and ages, and that human nature remains still the same, in its principles and operations,' and assertions such as, 'Would you know the sentiments, inclinations and cause of life of the Greeks and Romans? Study well the temper and actions of the French and English: You cannot be much mistaken in transferring to the former *most* of the observations which you have made with regard to the latter. Mankind are so much the same, in all times and places, that history informs us of nothing new or strange in this particular.'[217] The second charge levelled at Hume in relation to this part of our discussion is that despite his insistence on the uniformity of human nature, his own views are those of a particular culture at a particular point in time; he does not present us with humanity's virtues, but with those of an 18th century Edinburgh gentleman. This view is most clearly expressed by Alasdair MacIntyre, who says in *Whose Justice? Which Rationality?* that, 'What Hume presents as human nature as such turns out to be eighteenth-century English human nature, and indeed only one variant of that, even if the dominant one.'[218]

There is some justification to these criticisms. Hume evidently does expect to find a high degree of uniformity in human nature, and he does produce a catalogue of virtues

---

[214] *Truth and Truthfulness*, page 39.
[215] *Truth and Truthfulness*, page 42.
[216] For a recent discussion of this criticism of Hume, as well as a convincing reply to it, the line of which I shall largely follow here, see Simon Blackburn's book *Truth: A Guide for the Perplexed*, especially Chapter 8, Section 2: 'Mind Reading'. Interestingly, Blackburn includes Williams among those who have thought Hume incompatible with a genealogical approach, despite Williams' explicit use of Hume's account of justice as an example of imaginary genealogy in *Truth and Truthfulness* (pages 33-34).
[217] *Enquiry Concerning Human Understanding*, Section VIII, Part I, 65, page 83.
[218] *Whose Justice? Which Rationality?* page 295.

which root him firmly in a particular time and place, to the extent that some of the entries in this catalogue (such as chastity) appear almost quaint from a modern perspective, and his arguments for them (in the case of chastity, the assurance for a father of the paternity of his children) may even seem repellent.[219] However, these supposed failings are not fatal to his overall understanding of virtue, and should not necessarily be regarded as failings at all. Any account of morality or practical reason which makes use of the concept of virtue, and possibly any account of morality or practical reason at all, must suppose that there is *some* degree of uniformity in human nature. Only the most radical proponents of a blank-slate theory would claim that there is absolutely no such thing as human nature, and this is a claim which is at odds with our own experience, our knowledge of history and the behaviour and development of non-human animals. In particular, it is at odds with our everyday experience of the concepts labelled as virtues in Hume's and other virtue theories; part of the appeal of these theories is that they deal in concepts which are familiar to us, and at least some of these appear to be manifested and admired throughout human history. Faith that traits such as honesty, courage and benevolence are generally valued does not seem unwarranted.

Furthermore, despite his commitment to a common human nature, Hume both implicitly and explicitly allows for variations in the development and expression of this nature. This allowance is implicit in his distinction between natural and artificial virtues; this distinction allows that different cultures in different historical circumstances could exhibit and value character traits which are peculiar to them, yet are as deeply embedded within individuals and society to the degree that the respect for property Hume calls justice was embedded in his own. As mentioned earlier, Hume explicitly acknowledges the possibility of cultural variation in a dialogue included with the *Enquiries*. In this dialogue, Hume describes the superficially outlandish habits of two apparently fictional communities, before going on to reveal that these are actually societies which his readers would recognise: Periclean Athens and the France of Hume's own time. The point of this dialogue is eloquently expressed when Hume says that, 'The Rhine flows north, the Rhone south; yet both spring from the *same* mountain, and are also actuated, in their opposite directions, by the *same* principle of gravity. The different inclinations of the ground, on which they run, cause all the difference in their courses.'[220] In other words, Hume has an entirely plausible understanding of human nature; there are aspects of our nature which are common, and which produce common behaviour, values and judgements, and there are aspects of our nature which depend on material, historical and cultural circumstances, but which produce character traits which are nevertheless as central to our identities as anything produced by the common aspects of our nature. Given this understanding, the criticism of Hume's parochialism seems rather less relevant. Indeed, this understanding means that we should take Hume's apparent parochialism as a warning rather than a criticism. Despite the increased awareness in our own time of other cultures produced either by direct encounter or through education and the media we inevitably bring our own parochial concerns to the debate (and we must remember that Hume was, for his time, reasonably well travelled and conversant with other cultures). This means that it is particularly important to apply the Humean distinction which we are

---

[219] *Treatise*, Book III, Part II, Section XII, *Of chastity and modesty*.

[220] *Enquiries Concerning Human Understanding and Concerning the Principles of Morals, A Dialogue,* page 333.

attempting to apply here, between natural and artificial virtues, and in doing so must use methods such as imaginary genealogy to lift ourselves out of our parochial concerns, insofar as this is possible.

We can start our excursion into imaginary genealogy by imagining the same small society of human beings that Williams imagines in his State of Nature, with no writing and no elaborate technology. To this picture we can add that its inhabitants have basic human rational capacities and capabilities, and that they are faced with practical problems, the resolution of which is required for survival, to make the society more successful in material terms, or just to make its members more comfortable. Some of the practical problems faced by the society can be solved by agents deliberating and acting alone, while others require collective action and collective deliberation. We can include the presence of practical problems in the State of Nature without compromising it as we would do if we introduced other elements such as a particular form of government, or a particular set of environmental circumstances, because practical problems are always with us, whether they concern basic questions such as how to get enough food to avoid starving, or rather more exotic questions such as how to avoid missing a television programme. Any situation with a complete lack of practical problems would not be a recognisable situation. Even if we imagine a science fiction world in which all human material wants were satisfied, then we would still face dilemmas concerning relationships, power, social organisation and so on; a world in which all of these problems were solved without the need for human deliberation would be a horror story rather than science fiction. A world of practical problems, whether our world or the State of Nature, contains those pressures which produce Williams' virtues of Accuracy and Sincerity; the inhabitants of the State of Nature must be able to trust that those in a position to know more than them (such as the man who can see further because he has climbed a tree) are accurate in their beliefs and report those beliefs sincerely (when the man up the tree says that a lion is coming it is because he genuinely believes that a lion is coming and has a sound basis for that belief).

We can proceed in a similar step by step fashion from the pressures posed by the existence of practical problems within the State of Nature to the virtue of reasonableness.[221] The first step is one of simple survival. Human beings are best able to solve the most basic practical problems, such as how to get enough food to eat and how to shelter from the weather, through the application of their capacity for reason. Most animals do not address the practical problems which face them through the application of reason, or at least not in the same form as that exercised by human beings; their instinctive behaviour and physical capabilities provide them with a repertoire of means to address their particular range of needs. Humanity is not like that; nature may provide us with a set of instinctive responses and physical capabilities but we are, due to whatever evolutionary causes, so constituted that we must work things out for ourselves. Furthermore, our social natures, and the types of practical problem which face us (some of which are created by those social natures) mean that we cannot always work things out as individuals; sometimes we need to act collectively, and acting collectively usually means deliberating collectively. So, once we place human beings in a State of Nature

[221] In following these steps I am taking a similar path to that taken by Miranda Fricker for the virtue of epistemic justice in her forthcoming book *Epistemic Injustice*, particularly chapter 5, 'The Genealogy of Testimonial Justice'.

which contains practical problems, we create a need for those human beings to deliberate and act as individuals and in groups.

The second step towards the virtue of reasonableness is taken when we realise that the inhabitants of the State of Nature cannot achieve practical success through the application of the capacity for reason alone; such bare application would involve conducting deliberation in the face of every problem as if the agent had never encountered that problem or anything like it. The continued practical success of the inhabitants of the State of Nature demands that a repertoire of practical conclusions is developed in which individuals and groups can be confident; both so that those conclusions can be used again in similar situations, and so that they can be used as a basis for further deliberation.

The third step towards the virtue of reasonableness acknowledges that this confidence will sometimes be misplaced. Any individual or group will occasionally go wrong in practical deliberation, especially from such a basic starting point as we imagine in the State of Nature. This means that continued practical success is also dependent on the ability to critically appraise the contents of the repertoire of practical conclusions, even if these have worked before. At the same time, this critical appraisal must not be allowed to undermine confidence in the entire repertoire of practical conclusions; this would result in practical paralysis for the inhabitants of the State of Nature. These inhabitants must strike a balance between confidence and criticism.

The fourth step towards the virtue of reasonableness is necessary for the previous two steps to be effective; in order to be confident in the conclusions of deliberation, and in order to critically appraise those conclusions, we must be able to understand the deliberation which produced them; that deliberation must be intelligible. And this need for intelligibility produces the opportunity that constitutes our fifth and final step; if we understand our own deliberation and that of our potential partners in action we are in a position to build confidence in and critically appraise the way we deliberate as we would any other form of action. We can develop our repertoire of modes of deliberation as well as our repertoire of practical conclusions.

Although we have only considered each of these steps briefly, I believe that they are sufficiently plausible to be accepted as necessary consequences of a State of Nature filled with practical problems and populated by a basic human society facing those problems. These steps mean that the inhabitants of that society, if they value their own practical success, will value practical and deliberative behaviour which exercises the human capacity for reason to produce practical conclusions which are sustainable. In other words, they will recognise and value reasonableness. We can demonstrate this further by considering alternative ways in which societies and values might develop in the State of Nature, and ask whether these alternatives are at all plausible. Let us imagine three fictional communities. The first of these approximates to that which he have argued will arise within the State of Nature, and exhibits a respect for reasonableness, and we shall call it the *reasonable community*. The other two communities represent departures from this situation, and are manifestations of different types of unreasonableness. We shall call one the *lazy community* and the other the *pedantic community*. Before discussing them, though, we must be clear that these communities are not intended as basic or original versions of real human communities; they are wholly artificial. We do not imagine them in order to show what could arise in the State of Nature, but rather what

could not arise.

Members of the lazy community do not place a high value on the justification of practical conclusions and consequently are prepared to act on conclusions which have a shaky justification, or possibly no justification at all. Another way of putting this is to say that members of the lazy community have a high tolerance for arbitrariness. We can imagine a society tolerating arbitrariness to varying degrees, from allowing that practical conclusions based on completely arbitrary considerations are acceptable, to simply acknowledging that a degree of arbitrariness may be inevitable in all complex practical questions. At some point, of course, the degree of arbitrariness allowed becomes compatible with our virtue of reasonableness, and the lazy community which allows this degree of arbitrariness is indistinguishable from the reasonable community. The line between the two imaginary communities seems to be crossed when it is insisted that practical conclusions are at least justified on the basis of the their prospects of practical success; in other words, our imaginary lazy community accepts practical conclusions as good enough to act upon, even if there is no basis for believing that those conclusions will produce the practical results desired. There are, of course other grounds for the justification of action, but the fundamental question of practical success seems most appropriate to the State of Nature we are considering. And, as soon as we imagine a community which is prepared to accept the conclusions of practical deliberation even though they are not compatible with practical success we can see that such a community is either implausible from the outset or implausible as a sustainable prospect, depending on the degree of arbitrariness and lack of justification which we suppose it is willing to accept. If we imagine a complete tolerance for arbitrariness then we cannot imagine a community at all; all that we can imagine is a group of anarchic individuals acting on the basis of whatever whim or notion happened to cross their minds, and who would rapidly have to change their attitude towards justification or simply die out. If we simply imagine a heightened tolerance for practical conclusions which have no justification in terms of practical success, then we also end up with a community which will become extinct, albeit more gradually.

This extinction is likely to happen for two reasons. Firstly, a community with limited concern for the practical success of practical conclusions, to the extent that they will act in the absence of the prospect of that success will simply not be practically successful; at times its responses to practical problems will be fruitless, and this, in a situation such as we imagine the State of Nature to be, with no technology and limited resources, will eventually be fatal. We must be clear what we are talking about here: we are not simply saying that the lazy community sometimes goes wrong in practical matters; all societies and individuals make practical mistakes from time to time, sometimes trivial and sometimes catastrophic. Rather, we are saying that because the lazy community does not judge practical conclusions on the basis of whether they will produce practical success, it will not avoid mistakes, and may go on making a mistake even after it has been recognised as such. This brings us to the second reason that the lazy community will not last, even if it does not have a complete tolerance for arbitrariness; failure to judge practical conclusions according to their prospects of practical success means that the lazy community will grow neither its practical nor its deliberative repertoires. Because the lazy community essentially does not care about the quality of its practical conclusions, it will not develop the deliberative habits or establish the set of

deliberative starting points required for new rounds of deliberation to build on the success of the old. In the absence of a demand for justification, possible reasons for action stand on a par with each other; the prior success of a particular course of action or mode of deliberation provides no basis for preferring it over any other. This lack of development would not kill the lazy community on its own, of course, unless a particularly challenging problem happened along, perhaps in the form of a natural disaster or an aggressive predator; in the absence of such challenges a version of the lazy community that was not completely tolerant of arbitrariness could conceivably manage a desultory and somewhat odd existence. Extinction would come if the lazy community was ever forced to compete with a version of the reasonable community, which would not only be more successful in choosing fruitful courses of action, but would also be capable of incorporating that success within the criteria used for choosing future courses of action.

Of course, there is a third reason why the lazy community would become extinct: if inhabited by recognisable human beings it would either never come into existence in the first place, or it would change almost immediately into a version of the reasonable community. The need for practical conclusions to have some sort of correlation to practical success is so basic that we struggle to imagine how it could not form part of the criteria of judgement of reasons for action in every group of rational human beings; for this to be the case would require some very odd cultural circumstances lying far outside the imagined State of Nature. Surprisingly, though, circumstances have occurred historically with some of the characteristics of the lazy community: although they were short-lived, did not involve groups of people so cohesive that they could be described as a community, and did not pervade every aspect of deliberation for those people involved, there have been periods of delusion in which normal judgements of practical conclusions have been suspended, and a tolerance of arbitrariness has prevailed. Such periods are easiest to spot in the world of commerce and include phenomena such as the South Seas Bubble and the Dutch tulip mania of the 17th century, as well as a much more recent example: the dot com boom of the late 1990s. However, the existence of these anomalous periods need not trouble us too much, due to the unusual conditions required to sustain them, as well as the tendency for their collapses to be just as spectacular and rather more inevitable than their excesses.

The pedantic community is the opposite of the lazy community. While the inhabitants of the lazy community do not demand justification for practical conclusions to the degree required to reliably achieve practical success, the inhabitants of the pedantic community demand excessive justification for every practical conclusion. And, just as with the reasonable community, we can imagine varying degrees to which this defining characteristic of the pedantic community is exhibited by its inhabitants. We can start by imagining a version of the pedantic community whose inhabitants are never satisfied by any practical conclusion unless it can be shown to possess the certainty desired by our external reasons theorists, and diminish the degree of justification demanded until the pedantic community is indistinguishable from the reasonable community. The point at which we cross from pedantry to reasonableness mirrors the corresponding distinction between laziness and reasonableness. We said that we were imagining the lazy community when we imagined that its inhabitants did not at least demand that conclusions about action were made on the basis of their likelihood to produce practical success. We can similarly say that we are imagining the pedantic community when we

imagine a community whose inhabitants prize the identification of justified conclusions about action above actually taking action likely to produce practical success, and who would let opportunities for practical success pass if the justification for the action required to exploit those opportunities had not been established. Characterising the pedantic community in this way acknowledges that deliberation is itself an activity which takes time and effort – sometimes indefinite amounts of time and effort if a conclusion cannot be reached – and that sometimes going on deliberating is not compatible with successful practical action. The stereotypical inhabitant of the pedantic community is capable of being paralysed like Buridan's ass if offered two near-identical options for action and compelled to find finer and finer considerations to justify choosing one over the other.

Just as with the lazy community we can see that the most extreme example of the pedantic community would be a catastrophic failure; a community whose inhabitants sought fundamental justification for every action would simply fail to take any action at all. Furthermore, this would even be the case if, as desired by our external reasons theorists, there was some fundamental basis for action capable of providing external reasons. If there is such a basis then humans seem perennially incapable of clearly apprehending it; hence millennia of philosophical debate. Even if, say, Kant was right, then the inhabitants of the pedantic community would still be deliberating about whether he was right, just like the rest of us. However, again as with the lazy community, even milder forms of the excessive demand for justification which characterises the pedantic community seem to lead to the extinction of that community. In much the same way that the lazy community suffered from wasted effort and resources through pursuing insufficiently justified actions, the pedantic community would suffer from missed opportunities due to an inability to take decisions. Furthermore, the pedantic community is as unlikely to develop its deliberative and practical repertoire as the lazy community, albeit for different reasons. Because its inhabitants always demand excessive degrees of justification, the pedantic community is unable to develop confidence in its practical conclusions. Those conclusions which, however painfully, are established as justified for one set of circumstances may not be accepted as justified for a similar set of circumstances. Confidence implies that beyond a certain point justification exists without having to seek it, or even to articulate it. The inhabitants of the pedantic community cannot be content with such silence. Once again, under specialised circumstances the excessive demand for justification and consequent inability to develop confidence might not be fatal. However, such specialised circumstances are not found in the State of Nature and, just like the lazy community, we cannot imagine the pedantic community surviving an encounter with the reasonable community, either because it would be out-competed, or, more likely, because it would simply become the reasonable community.

However, we must remember that such a competition would not arise in the State of Nature for the pedantic community or the lazy community. Our inability to imagine the lazy community or the pedantic community surviving an encounter with the reasonable community does not indicate that they would be beaten in such encounters, or that the reasonable community is the best of three possible communities: it indicates that the reasonable community, which respects the virtue of reasonableness, is the only community we can imagine at all in the State of Nature. As a consequence, we can regard reasonableness as a natural virtue. This means that, not only can we expect to find

reasonableness exhibited in some form in all cultures, but also that, if we find it to be compatible with our account of internal reasons our common intuitions about reasons, it forms part of our understanding of practical reason, not just for those of us having this debate, but for all of humanity. We shall return to the question of whether the natural virtue of reasonableness is compatible with our account and our intuitions later. First, however, we should sound a note of caution about our identification of reasonableness as a natural virtue.

We should not imagine that we shall find reasonableness exhibited in exactly the same way in every culture; although we have said that the only community we can imagine in the State of Nature is the reasonable community, we should not suppose that humanity lives in uniformly reasonable communities, or that reasonableness is always manifested in the same way. To do so would be to deny the evidence of history. In a modern, Western, secular, 21st century democracy such as Britain, the idea of reasonableness has echoes of broadly liberal values, such as tolerance and fairness, as well as an uneasiness with dogma. However, we are also aware that other cultures, some current and some historical, do not share such values, and consequently that what they consider reasonable will often differ dramatically from what we consider reasonable. In understanding this variety we must recognise that the display of reasonableness has three aspects: the underlying motivations which lead us to deliberate and act in a reasonable fashion; the deliberative skills which enable those impulses to be expressed; and beliefs and judgements about what is sustainable, which vary from culture to culture and from person to person.

For example, consider the difference between a modern, Western, secular, liberal democrat and a devoutly religious Renaissance figure such as Thomas More. From what we know of More from his own writing and that of others we would certainly be inclined to describe him as reasonable, yet much of what he found to be reasonable, such as his own persecution of Protestants and his willingness to die rather than deny the authority of the Catholic Church, we would not find to be reasonable today. The difference between More and our modern liberal is not necessarily in their reasonableness, but in their starting points; they can find reasons which they individually consider sustainable, yet cannot follow the same deliberative routes because they do not start from the same place and do not have the same deliberative skills. This does not mean that reasonableness is a relativist notion, or that what different cultures consider to be reasonable and unreasonable is irreconcilable. On the contrary, the motivations and patterns of behaviour which constitute reasonableness remain the same, even if the context of the reasonable agent and the subject of his or her deliberations change. Furthermore, the nature of reasonableness itself gives us hope for connection between disparate cultures. I do not want to claim that reasonableness can solve all disputes or reconcile all differences between cultures and individuals; years or even centuries of engagement may be required for that, and may even entrench conflict. However, the impulses behind reasonableness push in the direction of reconciliation. The desire for reasons to be sustainable keeps us alive to the challenges arising from other points of view which indicate that our current reasons may no longer be sustainable.

3.3.2 Reasonableness and Internal Reasons

While we have been attempting to show that reasonableness is a virtue we have put our account of internal reasons to one side. We must now return to that account and ask whether the virtue of reasonableness complements it or contradicts it. We can ask this question at two levels: we can ask whether the individual elements of the internal reasons account are compatible with the concept of reasonableness as a virtue; and we can ask at a somewhat higher level whether the concept fills gaps within the account and extends our understanding of the dependence of practical reason on motivations. In answering this second question we must consider whether, if we establish it as part of the internal reasons account, the concept of reasonableness as a virtue allows that account to meet our common intuitions about reasons rather more completely.

We shall start by considering the five elements of the internal reasons account as they have developed in the course of our discussion. Let us briefly recap what these elements are. Firstly, the account rests on the basic claim that reasons are dependent on subjective motivations, and that consequently external reasons do not exist. Secondly, we maintain that reasons can be produced by a process of sound deliberation, and that this process can modify the agent's motivations. Sound deliberation is not restricted to algorithmic, stepwise, reasoning, although it may certainly contain such reasoning, but also includes less structured elements such as the exercise of the imagination. Thirdly, we insist that both an agent's reasons and what may count as sound deliberation are indeterminate. Fourthly, we claim that the mind itself contains indeterminate psychological contents, including the forerunners of beliefs and motivations, which are progressively settled into more determinate contents both by the formation of character over time and by individual instances of deliberation. Finally, we argue that even internal reasons capable of being produced by sound deliberation from the agent's motivations can be divided into those which are accessible to the agent, and those which the agent can only apprehend by overcoming some deliberative obstacle. In short, the internal reasons account denies any picture in which the agent's reasons are laid out clearly and cleanly irrespective of his or her identity, and in which the agent's task is to locate the set of reasons pertinent to current circumstances. Rather, it presents the agent's psychology and his or her reasons as developing and contingent entities, whose nature is not settled prior to the start of deliberation, and which will be at least partly determined by the way in which deliberation happens to go. It is this degree of contingency which simultaneously satisfies our individualistic intuition about reasons and troubles the universalistic intuition about reasons.

The concept of reasonableness as a virtue is compatible with the basic claim that reasons are dependent on motivations in two ways. Firstly, and most straightforwardly, unlike some of the external reasons theories we have encountered, the concept does not demand or imply that reasons exist in the absence of motivations. Secondly, and more importantly, it is compatible with the claim because it has been constructed on the basis of the implications of that claim. We have reached our understanding of the virtue of reasonableness by treating deliberation as an action and by asking what motivations might give us reasons to go about that action in a particular way, just as our account of internal reasons requires. Furthermore, the second and third order questions we have considered (what our attitudes are towards our motivations concerning reasons, and how our consciousness of these attitudes and desires for approval influence our behaviour) all have motivations at their root. It is, of course, unsurprising that an essentially Humean

claim about action should be compatible with an essentially Humean understanding of virtue.

The concept of reasonableness as a virtue is compatible with the understanding of sound deliberation within our account because it has the same influence on reasons and deliberation: it constrains the reasons which we judge to be legitimate without determining exactly what those reasons are and what that deliberation should be. Sound deliberation constrains reasoning because not just any way of deliberating can be considered sound, but does not determine reasons because there are many ways of deliberating soundly in any particular practical situation. Similarly, there are ways of deliberating and acting which are manifestly demonstrations of unreasonableness, but there are also many different ways of deliberating and acting in the same situation while remaining reasonable. Indeed, the relationships which sound deliberation and reasonableness have to our reasons resemble each other so closely that it is tempting to say that they are two terms for the same thing, or that one is a component of the other. I am reluctant to do this because there are distinct differences between the concepts, one of which is particularly important. Although we use the term reasonableness indiscriminately in everyday language to refer to reasons, actions and agreements, among other things, if we understand reasonableness as a virtue then we understand it primarily as a characteristic of agents. By contrast, sound deliberation is just what it sounds like: a description best applied to deliberation rather than agents. Because of this distinction it is possible for reasonable agents to deliberate unsoundly; an agent may be driven by all the motivations underpinning reasonableness, and may be sincerely attempting to deliberate in a reasonable manner, yet make some sort of error in deliberation. Just as we would continue to call someone honest who sincerely reported a false belief, we would continue to call someone reasonable who made an unwitting error which rendered an instance of deliberation unsound. Similarly, an agent may deliberate soundly yet be thoroughly unreasonable; the agent who is subject to the sly form of the vice of unreasonableness may follow an impeccably sound deliberative route which serves the purpose of distorting, suppressing or confusing the deliberation of others. So, I think that it is best for us to say simply that the concept of sound deliberation is compatible with the concept of reasonableness as a virtue, and that the reasonable agent will want to deliberate soundly, even if this desire is not always satisfied.

The way in which the concept of reasonableness as a virtue is compatible with the indeterminacy of reasons and deliberation is similar to the ways in which it is compatible with our understanding of sound deliberation. This is hardly surprising, as it is one of our claims that what counts as sound deliberation is essentially indeterminate. Our primary illustration of the implications of indeterminacy was to imagine near-identical agents in near-identical situations who can nevertheless follow different deliberative routes to radically different conclusions about their reasons for action, and to argue that these differing deliberative routes could be considered sound and the conclusions they issue in considered justified. Our understanding of reasonableness also allows that both agents could be considered reasonable. Of course, our understanding of reasonableness allows that different starting points for deliberation and different deliberative routes may be considered reasonable by different individuals and different cultures; but that is not the situation that we are considering here. Reasonableness as we understand it allows that different deliberative routes may be followed by near-identical agents because all it asks

is that the agent pursues sustainable reasons. And this allows that many different paths may be followed, even if these paths start from the same point and the agents pursuing them have much the same deliberative repertoire. However, as well as allowing for the situation described in our primary illustration of indeterminacy, the virtue of reasonableness also has a deeper relationship with indeterminacy. What may be considered reasonable is itself not fully determinate. Our definition of reasonableness, whether as the pursuit of sustainable reasons, or as the mean between laziness and pedantry in questions of practical reason, deliberately employs terms whose meaning is less than fully determinate. I believe that this essential indeterminacy is a necessary feature of a virtue understood in Humean terms, and potentially a virtue understood in any theoretical terms at all; although we may attempt to delineate and describe such virtues and even to speculate about their origins, when we do so we are building on an untheorised, unarticulated recognition of that virtue, without which we cannot genuinely understand it at all. The essential indeterminacy of reasonableness is such that we determine what is reasonable or unreasonable by the application of judgement and recognition to individual agents, reasons and actions, rather than through the application of rules or determinate criteria.

The concept of reasonableness as a virtue is also compatible with both of the senses in which we understand the phenomenon of steadying the mind. The sense we have concentrated on so far in our discussion is that in which the contents of the mind are settled within individual instances of deliberation; when thinking about what to do the agent is beset by a range of considerations and possibilities, some of which become settled into beliefs about what to do, and some of which become settled into desires, both for outcomes and for means of achieving those outcomes. The desire for sustainability underlying the virtue of reasonableness reflects this phenomenon; it expresses the wish that we should not only reach justified conclusions within particular instances of deliberation, but that those justified conclusions should become part of our deliberative furniture, providing an established start point from which the next instance of deliberation can proceed. Thus, steadying the mind, even in the context of individual instances of deliberation, does not just mean settling the mental contents required to carry out that piece of deliberation; it constitutes a development in the deliberative repertoire of the agent. Of course, this leads us to the second sense of steadying the mind, which has not concerned us so much in this discussion: steadying the mind as the establishment of character. As a virtue, the concept of reasonableness is thoroughly compatible with this sense of steadying the mind. Virtues are attainments of character which we expect to be realised through a long process of education, cultural immersion, personal experience and practice.[222] Each of these processes, along with the development of a deliberative repertoire through individual instances of deliberation, steady the unsettled contents of the mind into constituents of character. Furthermore, reasonableness reinforces some of the points Williams was trying to make through his discussion of steadying the mind. The fictional character of Rameau was interesting because the way in which his fleeting

[222] We have only briefly considered the means by which virtues are attained, partly through our discussion of the difference between virtues and skills, and partly through our excursion into imaginary geneaology, although the latter was of course not intended to show the actual attainment of virtue by actual agents. For fuller treatments of this subject from a largely Aristotelian perspective, see Sabina Lovibond's book *Ethical Formation* and M.F.Burnyeat's paper 'Aristotle on Learning to be Good'.

character was formed resembled the way in which all of our characters are at least partly formed (through the social situation he found himself in and the way in which he interacted with others) but the steadiness of his character did not resemble ours at all (it was continuously reconstituted as he flitted from one social situation to another) with the result that he was thoroughly unreliable, not through any maliciousness, but through sheer volatility. By contrast one of the effects of our mind becoming more steady for most of us is that we become more reliable, and this is certainly the case as the virtue of reasonableness becomes embedded in our characters.

Finally, we come to the significance and treatment of deliberative obstacles within our account. We said that such obstacles both create reasons which would otherwise not exist and obscure reasons which agents would otherwise accept, and I labelled those reasons to which the agent could deliberate soundly without having to overcome significant deliberative obstacles *strongly internal reasons,* and those reasons to which the agent could only deliberate if deliberative obstacles were overcome *weakly internal reasons*. The concept of reasonableness as a virtue not only fits well with this concept, but also offers us an alternative to this perhaps rather inelegant terminology, as the boundary of the reasons which may be acknowledged by the reasonable agent will often coincide with the boundary between strongly internal reasons and weakly internal reasons. The distinction between strongly and weakly internal reasons allows that the deliberation of agents can be considered sound even if it settles on reasons which are not as good as those which lie on the other side of deliberative obstacles. The importance of reasonableness to this idea is two-fold. Firstly, it seems natural to say that an agent who seeks and acts on reasons that can be apprehended through sound deliberation is reasonable, even if the path of that sound deliberation has been influenced by deliberative obstacles; although we must also acknowledge that sometimes the reasonable agent will be expected to overcome deliberative obstacles as part of being reasonable. Secondly, it also seems natural to say that an agent's perception of what is reasonable, formed by his or her historical and personal circumstances, can itself constitute a deliberative obstacle. When we originally discussed deliberative obstacles we considered several of the forms which such obstacles could take. One of those forms was deliberative habits; the tendency to follow the similar paths to those we have followed in the past, particularly if those paths have been successful. The virtue of reasonableness is not exactly the same as habit; habit has connotations of laziness and complacency which do not fit with the demand for sustainability. However, the demand for sustainability also means that we seek reasons which are good enough rather than reasons which are perfect, and the very discovery of reasons which are good enough and the continued practical success of action on those reasons may prevent us from apprehending reasons which are even better, or which are more suited to a changed situation.

For example, consider the Imperial Roman court as described by Tacitus. For the inhabitants of this particular historical situation it is entirely reasonable to take such practical precautions as having food examined for tampering, or taking bodyguards when visiting supposed friends. If you or I were thrust into such a situation without preparation we would find these precautions not just unreasonable but melodramatic, and as a consequence might not survive for long. Conversely, agents who had existed in a state of reasonable paranoia who found themselves in a new position where most people could be trusted most of the time would still tend to deliberate, on the basis of their prior

experience and knowledge, in a suspicious and fearful fashion. Given their background, such modes of deliberation would be manifestations of reasonableness, but this reasonableness would stand in the way of more trustful and fulfilling relationships with others.[223] This is why, as we argued earlier, reasonableness can sometimes be heroic; the imagination required to perceive that what seemed sustainable is no longer sustainable, and the effort required to get others to see that this is the case, may be extreme.

So, it seems that the concept of reasonableness as a virtue is at least compatible with the individual components of our internal reasons account. However, I believe that the concept does more for our account that simply provide us with a way of understanding the consequences of our common motivations and reasons; it provides us with a way of satisfying the universalistic intuition regarding reasons which we have so far been only partly able to satisfy. Remember that we found two ways to satisfy this intuition within the bounds of our account, even before introducing the concept of the virtue of reasonableness. Deliberation does not always go well. Sometimes it is beset by deliberative obstacles, meaning that while the agent can deliberate soundly to one set of reasons, an observer who is not subject to these obstacles may deliberate soundly to reasons which are both different and better. The observer's judgement may legitimately differ from the agent's, even though we also say that the agent has found genuine reasons. And sometimes, of course, the agent's deliberation is simply unsound; even without the help of deliberative obstacles, the agent has made some error of inference or logic and consequently supposes him or herself to have discovered reasons which do not actually exist. However, these two ways in which our account allows that observers can legitimately make better judgements about the reasons of agents than agents themselves do not do enough to satisfy the universalistic intuition. These two ways of going wrong in deliberation seem merely to be deficiencies in information or deliberative skill, which could be corrected with the help of the observer. Of course, it is not quite that easy; the deliberative obstacles we have discussed include habits and a lack of deliberative capabilities, which might take years of training and experience to change, if, indeed, they could be changed at all. However, the thought remains that these habits could be changed and capabilities established in principle, even if this would be practically difficult; there is a sense that deliberators who have gone wrong in the ways we have discussed are simply in want of the right sort of help, and that with the right help their perceptions of their reasons would happily fall in line with those of the observer.

The true concern of the universalistic intuition is the intractable allegiance or resistance of agents to reasons despite our attempts to persuade them otherwise. The universalistic intuition indicates that if we continue to disagree about reasons when deliberative obstacles have been cleared out of the way and unsound deliberation has been acknowledged or corrected, then we must have recourse to something that will settle the disagreement; there must be a way of showing that one of us is wrong, and that if whichever one of us is wrong persists in our error, we are doing something wrong. In

---

[223] Of course, maintaining this mode of deliberation would not be likely to be very reasonable for very long. For an agent to continue to stay generally suspicious in the face of trustworthy, honest and open behaviour from all others and to be considered reasonable he or she would have to have been subjected to an extreme betrayal. Unfortunately, such situations are not inconceivable. Jews and other groups who considered themselves to be accepted members of German society in the 1930s were subjected to a betrayal that would make it reasonable for them to doubt the intentions and attitudes of those they have lived alongside with apparently normal relationships for many years.

other words, we want our judgements to have a normative *bite* which goes beyond the correction of deliberative error. At the same time we must remember that, most of the time, our universalistic intuition about reasons is no more than an intuition; it is not an articulated principle, and it is not absolute. As an intuition, it can be satisfied with less than proof of the existence of universal, inescapable reasons; it can be satisfied with the demonstration that we have grounds for challenging the reasons of others on the basis of more than simply skill or circumstance. I believe that the concept of reasonableness as a virtue gives us these grounds.

Of course, reasonableness cannot satisfy the universalistic intuition by providing universal reasons; our understanding of reasonableness is such that agents can be considered reasonable even though they acknowledge and act for divergent and potentially contradictory reasons. However, it does provide a normative framework for judging the way agents come by their reasons and maintain faith in them; those people whose deliberation, reasons and actions possess certain characteristics are taken to have attained a virtue; while those whose deliberation, reasons and actions do not are taken to be in the grip of a vice. When we challenge the sustainability of an agent's reasons, we expect that, if that agent is reasonable, he or she will care enough about such challenges to respond in some way, just as an honest agent would feel compelled to respond to an accusation of lying. The reasonable agent whose reasons have been challenged is likely to attempt to find further justification for those reasons, or, if that justification cannot be found and the challenge is serious, to modify those reasons, or even to come around to the challenger's point of view. The important thing is that the challenged agent's reasonableness provides a point of engagement; unless the challenge is obviously frivolous or without merit, the reasonable agent cannot leave it unaddressed without feeling uncomfortable. This point of engagement does not provide us, as challengers and judges of the reasons of others, with guarantees that we are right, or that the agent will respond in the way we want, by accepting our judgements and challenges; but, as mentioned, I do not think that such guaranteed acceptance is what we are after. Rather, we are after a way in which our judgements can get a normative and motivational grip on the agent, and this is what the virtue of reasonableness gives us. The challenged agent will not necessarily come to agree with us, but if the agent is reasonable and our challenge is worthwhile, the agent cannot simply ignore us.

Despite the hold on the agent provided by the virtue of reasonableness, it may seem that it still fails to satisfy the universalistic intuition in two ways. To start with, it may seem that the grip provided by judgements of reasonableness is not strong enough or does not go deep enough; while it may get agents to engage with us about their reasons it does not determine what those reasons are. It gets us argument rather than obligation. I think that to criticise the virtue of reasonableness on these grounds, or to attempt to look for something further which might give us obligation is to make two mistakes. Firstly, it is to misunderstand the universalistic intuition. As we have argued, this intuition is not that we must have absolute, universal reasons with which we must all comply, regardless of our motivations or deliberative circumstances; as we understand it, it is that we expect that there must be some normative pressure that we can bring to bear on those people who persist on disagreeing with us about their reasons. We do not expect that the source of this pressure must determine reasons, that it guarantees agreement, or that it even guarantees that we are correct in our judgements about the agent's reasons. A desire for

or an insistence on the existence of absolute, universal reasons has moved beyond the universalistic intuition about reasons that most of us share to a specific type of theoretical position; a theoretical position which we have considered and rejected. The second mistake made by anyone claiming that the virtue of reasonableness does not do enough work to determine reasons is to suppose that reasonableness or any other virtue must precisely prescribe behaviour. Reasonableness does not determine reasons because it does not comprise a set of rules about which reasons can be considered legitimate or what forms of deliberation can be considered sound; rather, it is the expression of common attitudes to deliberation and reasons, and these attitudes influence the way in which agents go about finding and keeping reasons rather than determining what those reasons are. We do not judge people to be reasonable solely on the reasons for which they act, but rather on the means by which they come by those reasons, and the way in which they defend or modify them if challenged. Of course, some reasons are so obvious that any reasonable agent will discover and act for them; but only some reasons.

The relationship of reasonableness to reasons can be considered through analogy with the relationship between other virtues and the behaviour they produce, particularly with the relationship between the virtue of honesty and the expression of the truth. Attainment of the virtue of honesty does not mean that the honest agent will always discover the truth, and certainly does not determine what the agent believes to be true; rather, it entails a certain attitude towards the truth and towards expression of the truth. We must be careful to remember that this is only an analogy, of course. Without getting into questions concerning knowledge and the nature of truth which I am not well equipped to handle, we may suppose that the truth has a rather more determinate and independent existence than the reasons of a particular agent at a particular time. However, as long as we remember that it is merely an analogy, it is a useful one; honesty determines attitudes towards the truth rather than the truth itself; and reasonableness determines attitudes to reasons and deliberation rather than reasons themselves. The analogy with honesty is also helpful because it reminds us just how much purchase judgements of virtue can gain on agents. Judgements of honesty matter greatly to most agents; an accusation of dishonesty will almost always elicit a response, and often a vehement one. Furthermore, the implication that a judgement of character is being made is often drawn from the expression of a judgement about a particular claim; 'Are you calling me a liar?' is as recognisable a response to an accusation of dishonesty as, 'Are you saying that I'm lying?' Accusations of unreasonableness may not produce such a dramatic response, but they do sting and they do typically produce some form of response; and, after all, we have acknowledged that reasonableness is usually among the milder of the virtues.

The second source of suspicion that the concept of reasonableness as a virtue is inadequate to satisfy our universalistic intuition about reasons resembles the first, and is based on a general criticism of accounts such as ours which argue that reasons are dependent on motivations. Such accounts are open to the possibility that agents exist whose motivations give them reasons which are incomprehensible to us, and it is therefore possible that agents exist whose motivations do not incline them towards reasonableness. They are unreasonable versions of the sensible knave. When we challenge their reasons they may be indifferent to the grounds of our challenge, and to the existence of the challenge at all; reasonableness does not provide us with a point of

engagement for such agents. The fear is that these agents are not only difficult or impossible to persuade, but that they escape judgement altogether. This fear is founded on a genuine possibility. We can imagine two types of agent who lack the impulses which incline most of us to be reasonable. The first is admittedly rather implausible; the agent who does not care at all about the sustainability of reasons. This agent would be undismayed even if unable to explain the reasons on which he or she was acting to him or herself. However, within the terms of our account, such an agent would still have reason to behave reasonably, even if not directly inclined to reasonableness because, as we have seen, some degree of deliberation in accordance with reasonableness is required to be consistently practically successful. Assuming that the agent had some motivations, then he or she would therefore have indirect reasons to deliberate and act as if reasonable. The worry, therefore, is that such an agent only behaves as if reasonable for instrumental purposes, in much the same way that an agent who cared nothing for the truth could decide that being outwardly honest was more instrumentally successful than trying to maintain a network of deceit. Such an agent might respond to accusations of unreasonableness, because they indicated that his or her actions might be ineffective in realising their intended ends, but still would not care about being reasonable.

The second case is a character who is rather more plausible and rather more sinister; the agent who is concerned to make his reasons intelligible and sustainable to him or herself, but who does not care about the opinion, trust or judgements of others, and who therefore is not concerned to make his or her reasons outwardly intelligible or sustainable to others. We have already considered such a possibility when we considered the forms which the vice of unreasonableness can take, and identified the sly form of unreasonableness, in which the agent uses the forms of reasonable discourse in order to subvert the reasons and actions of others, while masking his or her own reasons. And the correspondence of the agent who does not share our motivations concerning reasons with a vice indicates how we can allay the fear that such an agent or the other variant we have just considered escapes judgement. Any virtue theory, whether it is rooted in motivations or otherwise, allows that people are not always virtuous; if they were, then virtue would not be considered an attainment and would not elicit praise. However, this does not mean that vicious people escape judgement, even if they are so vicious as to be incapable of understanding that judgement. We hope that they will respond to the expression of judgements about their behaviour, and that even if they do not yet grasp a virtue sufficiently to be motivated by perception of its lack, they care enough about the good opinion of others to change their ways. But, of course, if they do not care, this does not by itself invalidate our judgement, and does not prevent us from expressing our judgement in other, more practical ways, such as withdrawal of trust, denial of cooperation, ostracism and, depending on how the agent's viciousness manifests itself, all of the other less pleasant sanctions available to society. So, far from allowing wayward agents to escape judgement, the understanding of reasonableness as a virtue gives us a language in which judgement can be passed, even if the subject of that judgement is deaf to it.

To summarise, the claim that reasonableness is a virtue does not satisfy the universalistic intuition by giving us universal reasons which none of us can escape. Rather, it identifies a character trait which makes us sensitive to external judgements about our reasons, and also inclines us to seek reasons which can withstand the

judgements of others. This does not mean that people will always acknowledge the reasons which we want them to, or that we can judge them to be irrational if they do not; but it does mean that they should respond to warranted judgements about their reasons, and that we can judge them to be unreasonable if they do not. And I believe that this is all that the universalistic intuition requires; we may hope that we are always right in our judgements about our own reasons and the reasons of others, but we also accept that we can sometimes be wrong, and that consequently that what we should reasonably expect in response to our judgements is engagement rather than obedience. The concept of reasonableness as a virtue, then, is not only compatible with our account of internal reasons; it is essential to it as a means of understanding why we deliberate as we do, how we can judge the deliberation and reasons of ourselves and others, and how we can better satisfy the universalistic intuition even in the absence of universal reasons.

3.3.3 Reasonableness and External Reasons

We will conclude our discussion of reasonableness by asking whether the external reasons theorists we have considered not only have flawed arguments, but whether they are being unreasonable by promoting and defending them. Of course, it is tempting to claim that such people are unreasonable, if only for polemical reasons. There are aspects of unreasonableness which they are definitely not subject to: they have no tolerance for arbitrariness, or for the hidden influence of motivations. However, superficially they may seem to offend against reasonableness in two ways. Firstly, it may seem that, because we have identified flaws in their theories, that they are clinging to reasons which have become unsustainable. However, while some people who make apparent external reasons statements may be guilty of this, I do not think that this is the case for any of the external reasons theorists we have considered here. By contrast, while I hope that I have taken some steps in this thesis to show what is wrong with a particular set of external reasons theories, I would not pretend to have undermined them to the extent that continuing to defend them would be an example of unreasonableness; at the most I may have given their proponents another set of challenges to answer. The reasons on which the external reasons theorists we have considered base their arguments are not held unreflectively; the external reasons theorists have an extensive theoretical and motivational basis for holding them, and many possible means of response to challenge. Convincing these theorists that their reasons for arguing as they do are unsustainable might be an eventual future achievement, but for the time being all we have done is taken another step in a debate.

Secondly, it may seem that the deliberation of our external reasons theorists is pedantic, as it insists on the relentless exercise of pure reason to the extent that it sweeps away all foundations except those that can be found within reason itself. In order to save themselves from scepticism they carry on deliberating until they take themselves to have found a basis for reasons within some universal location, such as the structure of rationality, or the rational implications of inescapable elements of human nature. However, as far as we have found, we do not need to be rescued from scepticism through the discovery of such universal foundations; all we need are reasons that are sustainable. And if they are not good enough we go on deliberating until they are; but not to the extent of attempting to find an answer to all possible practical questions. The relentless reflection which characterises the external reasons theories which we have considered is

not, in the first instance, intelligible to us, not necessarily because we cannot understand the paths which are followed, but because we cannot understand why anybody would follow them.

However, there is a sense in which it is too harsh to describe external reasons theorists as unreasonable on this basis. They are not seeking the answers to questions about mundane practical problems, but to questions of extreme seriousness, whether they are considering particular moral dilemmas to which we appear to have no solutions, or are considering the general question of what could possibly be the basis for all reasons. So, given the seriousness of these questions, a degree of reflection so great that it undermines the sustainability of any normally satisfactory reasons may well be appropriate. What is not appropriate, however, is to suppose that settling such questions is necessary to settle any practical question at all; the seriousness of the question of what could provide the answer to all practical questions does not mean that all practical questions share in this seriousness. Consequently, we can say that external reasons theorists are reasonable in their approach to ultimate questions, even if we claim that they find the wrong answers, but unreasonable if this approach is taken to all instances of practical deliberation. In passing we should also note that this conclusion about external reasons theorists (that those we have considered should generally be judged reasonable even if their theories are wrong) illustrates the additional dimension of judgement which the concept of the virtue of reasonableness makes available to us. We can take it that external reasons theorists are motivated to seek the truth, and therefore that if their theories are incorrect then they have reasons to abandon them and seek alternatives. However, they are operating within a context which makes even their mistaken deliberation reasonable.

3.4 Consequences

So, we have done more than defend our account of internal reasons; through the introduction of the virtue of reasonableness, we have developed it to the point where it is capable of satisfying both the individualistic and the universalistic intuitions about reasons. Within our account, reasons belong intimately to the agent because they are dependent on the agent's material circumstances, deliberative circumstances and motivations. Reasons are not just peculiar to individual agents but to the deliberative paths followed by individual agents. Yet at the same time we have a normative framework for judging the reasons and deliberation of agents, and an expectation that agents will respond to our judgements at least by re-examining their reasons, and by engaging with us if their judgements continue to differ. Our position does not hold out the prospect that we can all agree on our reasons because those reasons are all the same; rather, it holds out the possibility that if we were all reasonable we could at least attempt to agree on the individual reasons that individual agents possess. Furthermore, our normative framework gives us a way to judge those people who not only have bad reasons, but also refuse to respond to our reasons or to engage with us. There is, of course, no further guarantee that this judgement of unreasonableness will have any influence on the unreasonable agent, but this does not invalidate the judgement, any more than persistent cowardice or dishonesty invalidate judgements about courage or honesty.

As we said at the outset of our discussion, the account we have ended up with is neither complete nor systematic; certainly not in the same way as the external reasons theories we have considered. Rather, it is a patchwork of claims and concepts, some descriptive and some normative, which are compatible with our experience of reasons and with the discipline we demand of a philosophical theory. However, I believe that such a patchwork is appropriate to our subject which, as we have seen, is in large part indeterminate and resistant to theory. It is tempting to try to draw our various claims and concepts together in to a systematic and unified whole, underpinned by some transcendent principle. But, as we have seen, this temptation may lead us to accounts which are satisfying as theory, but which are thoroughly at odds with our experience. Better, I think, to take the more humble approach of filling in the holes in our patchwork, paying sufficient respect to the elements of our account in which we already have confidence, and to our everyday understanding of our experience. This way we get an account which, like our reasons, is good enough.

Humility about the ambitions and scope of theory does not prevent it from having consequences for our everyday lives. Hume apparently set out to describe and explain morals and action rather than to prescribe them, but it would be disingenuous to suppose that he intended to leave everything as it was. And, while much of our discussion has been intended to establish a recognisable and resilient account of practical reason with predominantly theoretical implications, the introduction of the virtue of reasonableness has consequences for our everyday relationship with reasons. The adoption of any virtue theory necessarily has consequences, most importantly for moral development, either as part of upbringing or as part of the way we live our lives. Of course, because virtues are recognisable character traits, they are prominent in our moral development already; if they were not then it would be implausible to claim that they were identifiable as virtues. However, there is a different inflection of thought and judgement which accompanies the

recognition of a virtue as a virtue. Once we recognise a particular virtue we begin to realise that behaviour which contradicts that virtue is not just an isolated aberration; it is an expression of who we are and an influence on who we will be. The same is true of reasonableness. If we understand this virtue then when we are tempted to browbeat agents into acting for reasons which they don't accept, or to settle for reasons which are not good enough, or to subvert the reasoning of others for our own ends, or to issue edicts about reasons that brook no challenge, or to pursue justification far beyond what is called for, then we realise that giving way to this temptation would not just mar a particular instance of deliberation, but would be a step towards unreasonableness. We realise that we should stop bluffing or being lazy, sly, autocratic or pedantic, especially in philosophy; and start being reasonable.

Bibliography

J.E.J. Altham and Ross Harrison (ed.), *World, Mind and Ethics – Essays on the ethical philosophy of Bernard Williams*, Cambridge University Press, 1995

G.E.M. Anscombe, 'Modern Moral Philosophy' in Geach and Gormally, *Human Life, Action and Ethics: Essays by G.E.M. Anscombe*

Aristotle, *Ethics,* trans. J.A.K. Thomson, Penguin, 1955

Simon Blackburn, *Ruling Passions,* Oxford University Press, 1998

– *Truth: A Guide for the Perplexed*, Allen Lane, 2005

Michael Brady and Duncan Pritchard (ed.), *Moral and Epistemic Virtues*, Blackwell, 2003

David O. Brink, 'Moral Motivation' in *Ethics*, October 1997

M. F. Burnyeat, 'Aristotle on Learning to Be Good' in Rorty, *Essays on Aristotle's Ethics*

David Copp, 'Belief, Reason and Motivation: Michael Smith's *The Moral Problem*' in *Ethics*, October 1997

Edward Conze, *A Short History of Buddhism*, Unwin, 1980

Edward Craig, *Knowledge and the State of Nature: An Essay in Conceptual Synthesis*, Oxford University Press, 1990

Garret Cullity and Berys Gaut (ed.), *Ethics and Practical Reason*, Oxford University Press, 1997

Donald Davidson, *Essays on Actions and Events*, Oxford University Press, 2001

– 'Actions, Reasons and Causes' in Davidson, *Essays on Actions and Events*

John Dewey, *How We Think*, Prometheus Books, 1991

Denis Diderot, *Rameau's Nephew and D'Alembert's Dream*, trans. Leonard Tancock, Penguin, 1966

W.D.Falk, *Ought, Reasons and Morality*, Cornell University Press, 1986

– ''Ought' and Motivation' in Falk, *Ought, Reasons and Morality*

Philippa Foot, *Natural Goodness*, Oxford University Press, 2001

Miranda Fricker, *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford University Press, forthcoming (2006), pre-published on ePrints.bbk.ac.uk

Mary Geach and Luke Gormally (ed.), *Human Life, Action and Ethics: Essays by G.E.M. Anscombe*, Imprint Academic, 2005

James Gleick, *Genius: Richard Feynman and Modern Physics,* Abacus, 1992

Mary J. Gregor (ed.), *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy*, Cambridge University Press, 1996

David Hume, *A Treatise of Human Nature*, ed. Ernest C. Mossner, Penguin, 1969

– *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, ed. L.A. Selby-Bigge, Oxford University Press, 1975

Rosalind Hursthouse, *On Virtue Ethics*, Oxford University Press, 1999

Susan James, *Passion and Action: The Emotions in Seventeenth Century Philosophy*, Oxford University Press, 1997

Immanuel Kant, *Critique of Pure Reason*, Everyman, 1934

- *Groundwork of the Metaphysics of Morals* in Gregor, *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy*
- *Critique of Practical Reason* in Gregor, *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy*
- *The Metaphysics of Morals* in Gregor, *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy*
- *On a Supposed Right to Lie from Philanthropy* in Gregor, *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy*

Christine Korsgaard, *The Sources of Normativity*, Cambridge University Press, 1996

- 'The Normativity of Instrumental Reason' in Cullity and Gaut, *Ethics and Practical Reason*
- *Creating the Kingdom of Ends*, Cambridge University Press, 1996
- 'Skepticism about practical reason' in Korsgaard, *Creating the Kingdom of Ends*

Primo Levi, *The Periodic Table*, Abacus, 1975

Sabina Lovibond, *Ethical Formation*, Harvard University Press, 2002

Alasdair MacIntyre, *After Virtue*, Duckworth, 1981

- *Whose Justice? Which Rationality?*, Duckworth, 1988

John McDowell, 'Might there be external reasons?' in Altham and Harrison, *World, Mind and Ethics*

- *Mind, Value and Reality*, Harvard University Press, 1998
- 'Virtue and Reason' in McDowell, *Mind, Value and Reality*
- 'Are Moral Requirements Hypothetical Imperatives?' in McDowell, *Mind, Value and Reality*

Thomas Nagel, *The Possibility of Altruism*, Princeton University Press, 1978

- *Mortal Questions*, Cambridge University Press, 1979
- 'Subjective and Objective' in Nagel, *Mortal Questions*
- *The View from Nowhere*, Oxford University Press, 1986
- *Equality and Partiality*, Oxford University Press, 1991
- *The Last Word*, Oxford University Press, 1997

Onora O'Neill, *Constructions of Reason*, Cambridge University Press, 1989

- 'Consistency in action' in O'Neill, *Constructions of Reason*
- 'Universal laws and ends-in-themselves' in O'Neill, *Constructions of Reason*
- 'Kant after virtue' in O'Neill, *Constructions of Reason*
- 'Reason and Autonomy in *Grundlegung III*' in O'Neill, *Constructions of Reason*
- *Bounds of Justice*, Cambridge University Press, 2000
- 'Principles, Practical Judgement and Institutions' in O'Neill, *Bounds of Justice*
- 'Autonomy: The Emperor's New Clothes,' the inaugural address to the joint session of the Aristotelian Society and the Mind Association, 2003

Jean-Jacques Rousseau, *The Confessions*, Wordsworth, 1996

Amélie Oksenberg Rorty (ed.), *Essays on Aristotle's Ethics*, University of California Press, 1980

Michael Smith, *The Moral Problem*, Blackwell, 1994

- 'In Defense of *The Moral Problem:* A Reply to Brink, Copp and Sayre-McCord' in *Ethics*, October 1997

Geoffrey Sayre-McCord, 'The Metaethical Problem' in *Ethics*, October 1997

Roger J. Sullivan, *Immanuel Kant's Moral Theory*, Cambridge University Press, 1989

J.J.C. Smart and Bernard Williams, *Utilitarianism For and Against*, Cambridge University Press, 1973

Tacitus, *The Annals of Imperial Rome*, trans. Michael Grant, Penguin, 1956

R. Jay Wallace, 'How to Argue About Practical Reason' in *Mind*, July 1990

Bernard Williams, *Moral Luck*, Cambridge University Press, 1981

- 'Internal and external reasons' in Williams, *Moral Luck*
- 'Moral Luck' in Williams, *Moral Luck*
- 'Persons, character and morality' in Williams, *Moral Luck*
- *Ethics and the Limits of Philosophy*, Fontana Press, 1985
- *Making Sense of Humanity*, Cambridge University Press, 1985
- 'Internal reasons and the obscurity of blame' in Williams, *Making Sense of Humanity*
- 'Replies' in Altham and Harrison, *World, Mind and Ethics*
- *Truth and Truthfulness*, Princeton University Press, 2002

Linda Zagzebski, *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*, Cambridge University Press, 1996